JACQUELINE UBER SILVA

TEXT MINING COM UMA APLICAÇÃO NA VALIDAÇÃO DOS REGISTROS DE OCORRÊNCIAS POLICIAIS NA REGIÃO DA GRANDE FLORIANÓPOLIS

FLORIANÓPOLIS – SC

2005

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO
EM CIÊNCIA DA COMPUTAÇÃO

Jacqueline Uber Silva

TEXT MINING COM UMA APLICAÇÃO NA VALIDAÇÃO DOS REGISTROS DE OCORRÊNCIAS POLICIAIS NA REGIÃO DA GRANDE FLORIANÓPOLIS

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Prof. Dr. Paulo José Ogliari

Florianópolis, Agosto 2005.

TEXT MINING COM UMA APLICAÇÃO NA VALIDAÇÃO DOS REGISTROS DE OCORRÊNCIAS POLICIAIS NA REGIÃO DA GRANDE FLORIANÓPOLIS

Jacqueline Uber Silva

Esta Dissertação foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistemas de Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Raul S. Wazlawick, Dr.

Coordenador do Curso de Pós-Graduação

Banca Examinadora

_____

Paulo José Ogliari, Dr. (orientador), UFSC

_____

Dalton Francisco de Andrade, PhD., UFSC

_____

Francisco José Espósito Aranha Filho, Dr., FGV

_____

Aran Bey Tcholakian Morales, Dr., UNISUL

“O vento nunca sopra
a favor
de quem não sabe
para onde quer ir.”
(Sêneca)

Ao Sidnei é claro!

AGRADECIMENTOS

Agradeço ao meu marido Sidnei pelo amor, carinho e companheirismo em todos os momentos da minha vida, auxiliando-me sempre em tudo que precisei.

À minha família, pelo incentivo nos momentos difíceis, fazendo com que eu seguisse em frente. Agradecimento especial ao meu irmão José Lino Uber por ter ajudado-me sempre que precisei.

Ao meu orientador Dr. Paulo José Ogliari, pela oportunidade concedida a mim para poder fazer esse mestrado e pelo auxílio em todos os momentos.

Ao Doutor Francisco Aranha pelas valiosas sugestões que foram imprescindíveis para realização deste estudo.

Aos amigos Alberto Pereira de Jesus, Juliano Pacheco e Evanilde Maria Moser pela amizade e auxílio nos estudos.

À Diretoria de Combate ao Crime Organizado (DIRC) por estarem sempre à disposição e pelo fornecimento dos dados.

À Vera, secretária do PPGCC, pela atenção e eficiência nos serviços prestados.

A todos aqueles que contribuíram para a realização desse trabalho.

RESUMO

A Polícia Militar de Santa Catarina recebe diariamente uma média de 6.000 chamadas telefônicas. Estas chamadas geram registros de ocorrências (ROs). Para checar se o RO está classificado na natureza de operação correta, é necessário ler todos os ROs um a um, validando se a descrição da ocorrência está coerente com a natureza de operação informada. Essa tarefa torna-se árdua e pouco ágil, já que boa parte desses ROs não se encontram classificados corretamente. A mineração de texto auxilia neste processo, pois classificando e tratando esses documentos, facilita na procura e contagem dessas ocorrências policiais pela sua natureza de operação. A Secretaria de Estado da Segurança Pública e da Cidadania de Santa Catarina, que é o órgão responsável em contabilizar, fazer os levantamentos estatísticos e apontar diretrizes para o combate ao crime, necessita de um processo automatizado para validar os dados. Este estudo tem por objetivo minerar 2.684 ROs policiais do ano de 2003 na Região Metropolitana da Grande Florianópolis. Verifica com isso a confiabilidade e valida a classificação já atribuída ao RO, apontando registros com erro na classificação, que deverão ser revisados pelo analista da Diretoria de Combate ao Crime Organizado (DIRC). Com o *software* ABC *Clean*, pode-se corrigir automaticamente os erros de ortografia. Para a mineração dos dados, foi desenvolvido o *software* ABC *Mining* e utilizou-se o *Weka*® para a geração das regras de decisão do tipo Se..Então, utilizando como técnica de classificação, a árvore de decisão. Os *softwares* foram desenvolvidos na linguagem de programação *Delphi*®. Com a geração das regras de decisão, pode-se validar se o registro de ocorrência está classificado na natureza de operação correta e sugerir a DIRC que implemente em seu sistema de cadastro de RO essas regras, evitando-se assim que os novos RO sejam cadastrados com a natureza de operação distorcida em relação a descrição. Dos 2.684 RO estudados, 5,4396% deles encontravam-se cadastrados na natureza de operação incorreta. Deve-se levar em consideração que os dados enviados pela DIRC já foram previamente selecionados para atender algumas naturezas de operação, e já se esperava que muitos deles estivessem classificados corretamente.

Palavras-chave: *Text Mining*. Registro de Ocorrência. Recuperação da Informação. Classificação. Árvore de Decisão. Limpeza de dados.

ABSTRACT

Santa Catarina Military Police receives an average of 6.000 telephone calls every day. These calls generate occurrence reports (ROs). In order to check whether each occurrence report is classified under the correct nature of operation it is necessary to read all of them, one, by one, validating is coherently according to the informed nature of operation. This procedure becomes a rather difficult and not very practical task since some of these reports are not classified correctly. A procedure called text mining is a useful tool for this process, since classifying and dealing with these documents facilitates their search, counting and classification by the nature of their operation. The Secretary of Public Security and Citizenship in Santa Catarina, which is the agency in charge of accounting, carrying out statistical surveys and providing guidelines to prevent crime, needs an automatized process to validate the data. The present study aims at mining 2.684 police occurrence reports during the year 2003 in the metropolitan area of Florianópolis. The study also investigates their reliability and at the same time validates the classification already attributed to the occurrence report, pointing out error classification which must be reviewed by the board of directors of the Organization for Crime Prevention. With the ABC Clean software, spelling errors may be corrected automatically. For the mining of data, a software called ABC Mining was developed and the Weka® was used to generate the decision rules of the type "If … then ", using a classification technique called "the decision tree". The softwares were developed using Delphi® as the programming language. With the generation of decision rules, it is possible to validate if the occurrence report is classified under the correct nature of operation and suggest the board of directors of the Organization for Crime Prevention the implementation of theses rules in its occurrence report system, preventing new occurrences to be reported with the incorrect or distorted nature of operation, more specifically with regard to their description. Of the 2.684 reports analyzed, 5,4396% of them were classified under the incorrect nature of operation. The data sent by the board of directors of the Organization for Crime Prevention had previously been selected to meet some different operation natures and so it was expected that many of them had been correctly classified. Key-Word: Text Mining, Occurrence Report, Information Withdrawal, Classification, Decision Tree, Data Cleaning.

LISTA DE ABREVIATURAS

BOE - Batalhão de Operações Especiais

CIASC – Centro de Informática e Automação do Estado de Santa Catarina

COE - Companhia de Operações Especiais

CRISP-DM – Cross-Industry Standard Process for Data Mining

DBCT – Descoberta de Conhecimento em Bases de Dados

DCT – Descoberta de Conhecimento em Texto

DIRC – Diretoria de Combate ao Crime Organizado

EI - Extração de Informação para dados não estruturados

PMSC – Polícia Militar de Santa Catarina

RI –Recuperação da Informação

RO – Registro de Ocorrência

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, as empresas e órgãos públicos estão “alimentando” cada vez mais as suas bases de dados, tornando-as geralmente muito grandes ou complexas. Vários fatores têm contribuído para este comportamento; entre eles, a queda nos custos de armazenamento e a disponibilidade de computadores de alto desempenho a um custo baixo.

Tanto o sucesso como o fracasso das empresas e organizações está intimamente associado à sua capacidade de lidar com dados e informações, os quais podem gerar novos conhecimentos e rapidez para a tomada de decisões estratégicas.

Na área de segurança pública brasileira, os dados estão armazenados em diferentes locais e em diferentes formatos, incluindo textos não estruturados. Mesmo com o esforço realizado pelas autoridades em integrar esses dados, nota-se que esse processo ainda está incipiente, pois existem disparidades entre os estados brasileiros.

Tan (1999) afirma que 80% das informações de uma empresa estão em formato textual. Entretanto, as organizações têm dificuldade para tratar adequadamente a informação não estruturada. A área de *Text Mining* ou Descoberta de Conhecimento em Textos – DCT, surgiu para minimizar este problema, facilitando na exploração do conhecimento armazenado em meios textuais.

A descoberta de conhecimento em textos pode ser definida como o processo de extrair padrões ou conhecimento, interessantes e não-triviais, a partir de documentos textuais. Nessa dissertação foram minerados os Registros de Ocorrência (RO) da Secretaria de Segurança Pública e Cidadania de Santa Catarina. Um RO é um documento legal elaborado pela Polícia que representa a primeira notificação oficial de uma queixa-crime para a maior parte dos casos que são encaminhados a uma unidade de polícia judiciária. Considerando as especificidades das funções de polícia investigativa e de polícia judiciária, pode-se dizer que o RO expressa o atendimento preliminar oferecido ao público e agentes institucionais que, por diversas motivações, acionam os serviços policiais civis. Ao cadastrar-se um RO é atribuído a ele uma natureza de operação. A natureza de operação determina qual o delito foi cometido. Para esse estudo

foram utilizadas 17[1] categorias de natureza de operação dos 732 tipos existentes. Como parte desses ROs encontram-se classificados em natureza de operação incorreta, utilizou-se uma metodologia capaz de identificar os ROs classificados incorretamente e apontar a provável natureza de operação correta. Para isso compara-se a descrição do RO com a natureza de operação já atribuída utilizando o processo de classificação. A classificação consiste no processo de identificação da categoria que um RO pertence e é uma das tarefas mais referenciadas na literatura. É também denominada de aprendizado supervisionado, pois a entrada e a saída desejadas são fornecidas previamente. Por exemplo, pessoas podem ser previamente grupadas nas classificações de bebês, crianças, adolescentes, adultos e idosos. Pode-se tomar como exemplo o RO número 1030271, onde o policial informou a natureza de operação C903 – comunicação falsa, e a descrição do RO tem como palavras-chaves arma, projéteis, óbito que informa claramente que o RO pertence a natureza de operação D309 – óbito no local.

Quando uma ocorrência policial está cadastrada na natureza de operação incorreta, produz dados estatísticos distorcidos, que por sua vez, leva a polícia a elaborar políticas de combate ao crime nem sempre eficazes.

1.1 Problema da pesquisa

Diante dessas considerações, as seguintes questões de pesquisa podem ser levantadas:

a) como utilizar os dados cadastrados nos registros de ocorrência de forma correta?
b) como limpar as bases de dados contendo os registros de ocorrência da Secretaria de Estado da Segurança Pública e Defesa do Cidadão?
c) como corrigir as palavras com grafia incorreta?
d) como reclassificar um documento anteriormente classificado?
e) como se pode agir proativamente para evitar novas inconsistências de dados?

[1] Consulte o anexo A para visualizar os tipos de natureza de operação estudadas

f) quais os *softwares* que deverão ser implementados para solucionar o problema?

1.2 Objetivos

Os objetivos desse estudo sobre *Text Mining* aplicado na validação dos registros de ocorrência policiais na Região da Grande Florianópolis, estão detalhados nessa seção e foram utilizados como guia para a condução e desenvolvimento desta pesquisa.

1.2.1 Objetivo geral

Aplicar uma metodologia com base em uma descrição textual dos registros de ocorrência permitir a limpeza da base de dados e a reclassificação e classificação dos mesmos.

1.2.2 Objetivos específicos

Os objetivos específicos são:

a) definir e implementar uma estrutura de dados;
b) definir procedimentos de preparação e limpeza de dados;
c) implementar um *software* para a limpeza dos dados;
d) descrever os dados de forma sintética;
e) desenvolver as regras de decisão com a utilização do *software Weka*®;
f) aplicar as regras de classificação geradas pelo *Weka*®;
g) validar se a classificação atribuída ao registro de ocorrência está correta;
h) validar o modelo proposto, verificando se a solução encontrada é coerente.

1.3 Justificativa

É importante ter em mente que o trabalho científico deve contribuir para o avanço do conhecimento. A justificativa deve evidenciar a contribuição que a pesquisa trará à ciência, à comunidade e ao pesquisador (AZEVEDO, 1999).

Avaliando-se a base de dados fornecida pela Secretaria de Estado da Segurança Pública e Defesa do Cidadão de Santa Catarina constatou-se alguns fatores que contribuem para justificar a aplicação da mineração de textos:

a) o elevado número de erros de ortografia;
b) a ausência de um corretor que limpe a base de dados automaticamente;
c) a descrição da ocorrência não condiz com a natureza de operação informada.

Devido a grande quantidade de ROs apresentarem inúmeros erros de ortografia, palavras sem separação por espaço, como por exemplo “assaltoamãoarmada”, muitas abreviações como fem, vtr, bo tornou inviável a mineração dos textos sem inicialmente a padronização e limpeza da base para, com isso, corrigir todas essas inconsistências. Como não existe nenhum *software* que faça a limpeza da base automaticamente foi construído o ABC *Clean*. Para a padronização da base de dados formou-se um *layout* padrão.

Com a geração das regras de decisão, tanto os registros existentes podem ser reclassificados como os novos podem ser classificados no momento da sua inclusão. Hoje os dados estatísticos não revelam a realidade das ocorrências policiais, devido a inconsistência entre a natureza de operação e sua descrição.

A aplicação das técnicas de *text mining* possibilita identificar e classificar corretamente a natureza de operação dos registros de ocorrência, garantindo uma maior confiabilidade na análise dos dados. A vantagem da mineração é produzir um modelo de fácil interpretação para os especialistas em segurança pública.

1.4 Estrutura do trabalho

O primeiro capítulo, relaciona a introdução, os objetivos, justificativas e a estrutura do trabalho.

O segundo capítulo, aborda a fundamentação teórica sobre Descoberta de Conhecimento em Textos (DCT).

No terceiro capítulo, explana-se como são os dados e os métodos adotados para a solução do problema e definição de termos técnicos e específicos sobre Segurança Pública.

O quarto capítulo descreve as duas aplicações desenvolvidas e o *software* utilizado para geração da árvore de decisão.

O quinto capítulo traz as conclusões e sugestões para trabalhos futuros.

Finaliza-se com as referências, apêndices e anexos.

1.5 Limitações do trabalho

Os registros de ocorrência utilizados nessa dissertação estão limitados à Grande Região de Florianópolis no ano de 2003.

2 DESCOBERTA DE CONHECIMENTO EM TEXTOS

Atualmente, num mundo moderno e globalizado, as empresas e os órgãos públicos buscam absorver o conhecimento de forma rápida e segura. Neste sentido, observa-se que a maior parte de suas informações encontram-se de forma não estruturada.

Documentos digitalizados, e-mails, memorandos, registros de reclamações de clientes, registros de ocorrências, dentre muitos outros exemplos, não encontram-se em uma mesma base de dados. Desta forma, questiona-se como agrupar todas essas informações e gerar novos conhecimentos. Para isso, desenvolveu-se a DCT. Pode-se encontrar na literatura várias formas de mencionar a DCT: Mineração em Textos, Descoberta de Conhecimento em Textos ou "*Text Mining*".

Segundo Fayyad (apud SILVA, 2002), a descoberta de conhecimento ocorre por meio de complexas interações realizadas entre homem e uma base de dados, geralmente por meio de uma série heterogênea de ferramentas. Loh (apud WIVES, 2003), afirma que as três grandes áreas que lidam com informações em grandes bases de dados são: *Data Mining* (Mineração de Dados para dados estruturados – DCBD); *Information Extraction* (Extração de Informação para dados não estruturados – IE); e *Information Retrieval* (Recuperação da Informação – IR) para textos ou palavras.

Para Loh et al. (2003) a tecnologia de *Text Mining* serve para "identificar os conceitos presentes nos textos. Conceitos representam "entes" do mundo real (entidades, eventos, objetos, sentimentos) e permitem entender que temas estão presentes nos textos ou do que tratam os textos. Em seguida, a exploração é feita através de um processo automático de mineração. Nesta etapa, são aplicadas técnicas estatísticas sobre os conceitos extraídos dos textos livres, na etapa anterior. Esta mineração é feita analisando-se a distribuição dos conceitos em coleções (a freqüência ou probabilidade com que aparecem) e a relação dos conceitos entre si, para descobrir associações e dependências".

Existem duas formas de mineração em textos: a DCT e a DCBD. A única diferença entre as duas é que, na DCBD os dados encontram-se estruturados, e na DCT estão em forma de texto. A DCT é uma metodologia para a análise de textos, que

permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associações e regras para classificação, e realizar análises qualitativas ou quantitativas. Soluciona grande parte dos problemas relacionados à busca, recuperação e análise de informações. Sistemas de informação que ofereçam características de DCT podem beneficiar os empresários e órgãos públicos, auxiliando-os a coletar e analisar os dados necessários à tomada de decisão e permitindo com que se posicionem melhor em suas atribuições. Segundo CHEN et al. (2004), "os investigadores humanos com anos da experiência podem frequentemente analisar tendências do crime, mas conforme a incidência e a complexidade dos crimes aumentam, erros humanos ocorrem, o tempo de análise aumenta, e os criminosos têm mais tempo para destruir a evidência e escapar. Aumentando a eficiência e reduzindo erros, as técnicas de mineração de dados criminais podem facilitar o trabalho dos policiais e permitir que os investigadores dediquem seu tempo em outras tarefas".

2.1 Conceitos básicos

Para um melhor entendimento sobre *Text Mining* é necessário que se conheça alguns conceitos básicos auxiliando na assimilação da técnica de mineração em bases textuais. Assim, a presente sub-seção desse capítulo apresenta esses conceitos e as principais características dessa tecnologia.

2.1.1 Stopwords

Segundo Santos (2002), *stopwords* "são palavras que ocorrem freqüentemente em textos. Uma vez que elas são muito comuns, sua presença não contribui significativamente para a determinação do conteúdo do documento". Logo, elas podem ser removidas do documento, para fins de *Text Mining*. Exemplos disso: os artigos, preposições, pronomes e demais palavras utilizadas para auxiliar na construção sintática das orações.

As s*topwords* são palavras muito freqüentes na coleção de documentos como um todo, sendo de baixa discriminação e portanto inúteis para representar e distinguir os documentos uns dos outros.

Para Wives (1999), pode ser traduzida como “palavras negativas”, “palavra-ferramenta” ou “palavras vazias”. Sua remoção contribui para aumentar a rapidez de operação, pois uma busca que emprega o termo “de”, certamente, recupera quase todos os registros em uma base de dados.

A Figura 1 sugere uma lista padrão das palavras que normalmente se encaixam no conceito de s*topwords*. Caso exista uma palavra que se repita muito na base de dados e que não tenha importância no processo de busca, esta deve ser incluída na lista.

a	desligado	faz	ou
acerca	deve	fazer	outro
agora	devem	fazia	para
algumas	deverá	fez	parte
alguns	direita	fim	pegar
ali	diz	foi	pelo
ambos	dizer	foram	pessoas
antes	dois	horas	pode
ao	dos	iniciar	poderá
apontar	e	início	podia
aquela	é	ir	por
aquelas	ela	irá	porque
aquele	ele	isto	povo
aqueles	eles	ligado	primeiro
aqui	em	maioria	qual
atrás	enquanto	maiorias	qualquer
bem	então	mais	quando
bom	está	mas	quê
cada	estado	mesmo	quem
caminho	estão	meu	quieto
cima	estar	muito	saber
com	estará	muitos	são
como	este	não	sem
comprido	estes	nome	ser
conhecido	esteve	nós	seu
corrente	estive	nosso	somente
das	estivemos	novo	tal
debaixo	estiveram	o	também
dentro	eu	onde	tanto
desde	fará	os	tem

Figura 1 - Lista de *stopwords*

2.1.2 Corpus

É um conjunto de dados lingüísticos, organizados seguindo alguns critérios, de maneira que sejam representativos do conjunto de dados, armazenados de tal modo que possam ser processados por computador, com a finalidade de gerar resultados úteis para a descrição e análise (SARDINHA, 2004, p.9).

Segundo Sardinha (2004), existem dois tipos de *corpus*:

> a) um *corpus* de estudo, representado em uma lista de freqüência de palavras. O *corpus* de estudo é aquele que se pretende descrever;
> b) um *corpus* de referência, também formatado como uma lista de freqüência de palavras. Também é conhecido como "*corpus* de controle", e funciona como termo de comparação para a análise. A sua função é a de fornecer uma norma com a qual se fará a comparação das freqüências do *corpus* de estudo. A comparação é feita através de uma prova estatística selecionada pelo usuário (qui-quadrado ou log da Verossimilhança). As palavras cujas freqüências no corpus de estudo forem significativamente maiores segundo o resultado da prova estatística são consideradas chaves, e passam a compor uma listagem específica de palavras-chaves.

O *corpus* de referência não deve conter documentos relacionados ao tema a ser analisado, pois ao comparar-se o *corpus* de referência com o *corpus* de estudo, a diferença entre eles será uma lista de palavras-chaves ou *keywords*.

O *corpus* de referência utilizado nesta dissertação é composto por vários documentos que foram utilizados como fonte desta pesquisa. Na página 80, pode-se visualizar parte do *corpus* de referência.

2.1.3 Keywords

Keywords ou palavras-chave, são aquelas cujas freqüências são estatisticamente diferentes no *corpus* de estudo em relação ao *corpus* de referência. Segundo Sardinha (2004), "para se analisar quais são as *keywords* necessita-se de dois elementos básicos: *corpus* de estudo e *corpus* de referência".

As palavras cujas freqüências no *corpus* de estudo forem significativamente maiores segundo o resultado da prova estatística são consideradas chave, e passam a compor uma listagem específica de palavras-chave. Na página 82, pode-se visualizar as *keywords* encontradas pelo *software ABC Mining*.

Sardinha (2004) sugere o seguinte algoritmo para extração de palavras-chave:

> a) selecione o primeiro item na lista de palavras do *corpus* de estudo;
> b) procure por este item na lista de palavras do *corpus* de referência;
> c) se o item constar do corpus de referência, vá para o passo a seguir, senão passe para o passo g;
> d) compare as freqüências através de uma prova estatística escolhida pelo usuário (verossimilhança é 'default', mas qui-quadrado também está disponível);
> e) se o resultado da comparação for estatisticamente significativo (segundo o nível de significância definido pelo usuário), copie esta palavra para uma nova lista, e chame-a de lista de palavras-chave;
> f) repita este procedimento até o último item da lista de palavras do *corpus* de estudo;
> g) se um item constante da lista de palavras do *corpus* de estudo não aparecer na lista de palavras do *corpus* de referência, assuma freqüência 0 para o item no *corpus* de referência;
> h) execute os passos d, e, e f.

2.1.4 Collocations

Segundo Santos (2002), "as colocações ou expressões compostas são agrupamentos de palavras onde o significado do todo é a soma dos significados das partes mais algum componente semântico adicional não previsto pelas partes".

Manning & Schütze (1999, p.141), define *collocations* como "uma expressão que consiste em duas ou mais palavras que correspondem a algum modo convencional de dizer alguma coisa".

Choueka (apud MANNING & SCHÜTZE, 1999, p. 145), afirma:

> "Uma colocação é definida como uma seqüência de duas ou mais palavras consecutivas que têm características de uma unidade sintática e semântica, e cujo significado ou conotação exato e não-ambíguo não possa ser derivado diretamente a partir do significado ou conotação de seus componentes".

Para Firth (apud MANNING & SCHÜTZE, 1999, p. 141), "Colocações de uma dada palavra são afirmações dos lugares comuns ou habituais daquela palavra."

Um exemplo disso está na expressão "chutar o balde", não está falando sobre baldes, mas é uma forma de expressão utilizada no idioma português para expressar que não se está preocupado com o que vai acontecer depois de determinada atitude.

2.1.5 Stemming

Porter (1997), afirma que *"stemming* consiste em converter cada palavra para seu 'radical' ('*stem'*), isto é, uma forma neutra com respeito a *tag-of-speech*[2] e inflexões verbais plurais. Por exemplo, as palavras '*learning*' e '*learned'* são ambas convertidas para o *stem* '*learn'*. "

Segundo Chaves (2004, p. 2), *"stemming* consiste em reduzir todas as palavras ao mesmo *stem*[3], por meio da retirada dos afixos da palavra, permanecendo apenas a raiz dela".

O propósito, segundo Chaves (2004, p. 2) é:

> Chegar a um *stem* que captura uma palavra com generalidade suficiente para permitir um sucesso na combinação de caracteres, mas sem perder muito detalhe e precisão. Um exemplo típico de um *stem* é 'conect' que é o *stem* de 'conectar', 'conectado' e 'conectando'. Dois erros típicos que costumam ocorrer durante o processo de *stemming* são *overstemming* e *understemming. Overstemming* se dá quando a cadeia de caracteres removida não é um sufixo, mas parte do *stem*. Por exemplo, a palavra gramática, após ser processada por um *stemmer*, é transformada no *stem grama*. Neste caso, a cadeia de caracteres removida eliminou parte do *stem* correto, a saber 'gramát'. Já *understemming* ocorre quando um sufixo não é removido completamente. Por exemplo, quando a palavra 'referência' é transformada no *stem* 'referênc', ao invés do *stem* considerado correto 'refer'.

Em 1980, Martin Porter desenvolveu um algoritmo que tem por objetivo o tratamento de *stemming* para a língua inglesa. Tem sido adaptado para várias línguas latinas, tais como espanhol e português.

2.1.6 Limpeza dos dados *(Data Cleaning)*

Para construir bons modelos precisa-se de dados depurados. Entretanto, os dados de muitas organizações possuem baixa qualidade. Valores ausentes, valores ilegais, combinações inexistentes e erros de grafia podem alterar seus resultados. Os recursos de transformações e limpeza nos dados aumentam o valor desses dados.

---

[2] Modo de falar

[3] Conjunto de caracteres resultante de um procedimento de *stemming*

Louzada Neto & Diniz (2000, p.35) afirma que “a qualidade dos dados é essencial para a obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração”.

As atividades de obtenção e limpeza dos dados normalmente consomem mais da metade do tempo dedicado ao projeto. Porém a limpeza dos dados pode evitar que a consolidação dos dados sejam distorcidos.

Geralmente são erros pequenos e simples, onde uma letra é adicionada, trocada ou omitida. São erros difíceis de serem encontrados em um conjunto de dados pela pequena diferença na ortografia.

Segundo Han & Kamber (2001, p.109) as tarefas para limpeza da base são:

a) preencha os valores que estiverem faltando;
b) identifique *outliers* e retire os ruídos dos dados;
c) verifique a consistência dos dados;
d) resolva a redundância causada pela integração dos dados.

Para a limpeza da base de dados da Secretaria de Segurança Pública do Estado de Santa Catarina, desenvolveu-se o *software ABC Clean* com a ajuda do dicionário de palavras *Aspell*.

2.1.7 Aspell

O projeto GNU foi lançado em 1984 com o intuito de construir softwares de código aberto. Um dos softwares construídos foi o *Ispell* que é um programa que ajuda o usuário a corrigir a soletração e erros tipográficos de uma palavra. Quando encontra uma palavra que não esteja no dicionário, o *Ispell* tenta encontrar a mais próxima e exibe uma lista de sugestões de palavras, ou ainda, o usuário pode incluir a palavra que ele não encontrou. Permitindo ao *Ispell* aprender novas palavras (GNU, 2005).

O *Aspell* é um corretor ortográfico, de código aberto, desenvolvido para substituir o *Ispell*. Sua principal característica é que ele faz um trabalho de sugestão de palavras, que é feita através do critério de proximidade fonética, muito melhor que o *Ispell*. O *Aspell* é também uma biblioteca, todavia, a maneira recomendada para se usar o *Aspell* é

através da biblioteca *Pspell*, uma vez que a verdadeira interface da biblioteca do *Aspell* é constantemente modificada.

A idéia base consiste em identificar palavras de um texto como sendo corretas. Para ser correta, uma palavra deve estar presente numa lista de palavras corretas (dicionário) ou ser uma derivação de alguma palavra existente (plurais, feminino, etc). Caso contrário, determina-se um conjunto de palavras extraídas da lista que sejam semelhantes a errada. Esse conjunto é apresentado como alternativo a palavra errada. A semelhança é apenas determinada por regras lexicográficas, como o número de letras que se tem de modificar para a transformar na outra.

2.2 Algoritmo de Aprendizado de máquina

Consiste em um programa de computador que realiza um conjunto de atividades capaz de melhorar seu desempenho a cada experiência, em um processo de aquisição automática de conhecimento a partir de novos dados.

Segundo Rezende (2003, p.90) o aprendizado de máquina (AM) é:

> "uma área de pesquisa cujo objetivo é o desenvolvimento de sistemas computacionais capazes de aprenderem e adquirirem conhecimento de forma automática. Um algoritmo de aprendizado é um programa computacional capaz de tomar decisões baseadas em soluções de problemas anteriores. Os sistemas de aprendizado possuem características que permitem sua classificação de acordo com a linguagem de descrição, modo paradigma e forma de aprendizado. As pesquisas em AM buscam construir programas de computador que melhorem o seu desempenho em alguma tarefa por meio de experiência".

Recentemente, a interligação das áreas de AM e de Extração de Conhecimentos em Bases de Dados tem se tornado cada vez mais importante, na medida em que a manipulação e análise manuais deste grande volume de dados armazenados pelas aplicações têm se tornado inviáveis.

Pode-se dividir de forma geral os algoritmos de aprendizado, sejam métodos estatísticos, coneccionistas ou de AM em dois grupos distintos: aqueles que realizam aprendizado supervisionado e aqueles que realizam aprendizado não supervisionado.

2.2.1 Tipos de aprendizado de máquina: supervisionado e não supervisionado

Algumas ferramentas utilizam o aprendizado supervisionado e outras o não supervisionado (PRADO, 1997). Considera-se o aprendizado supervisionado quando o processo baseia-se em exemplos. Os exemplos e os resultados esperados são apresentados ao algoritmo de aprendizado. Esse tenta adaptar-se a fim de produzir o resultado esperado para cada exemplo. O processo é repetido até que o algoritmo apresente os resultados esperados com o mínimo de erro possível.

O aprendizado chamado de não supervisionado quando se possuem os dados, mas não se possuem modelos ou exemplos que possam ser ensinados ao algoritmo.

Nesse caso, o algoritmo fica encarregado de identificar alguma espécie de relacionamento entre os dados. Os relacionamentos identificados são apresentados a um especialista que deve então validar a relação e encontrar algum significado para ela.

2.2.2 Descoberta reativa e proativa

Existem dois modos de descoberta: reativa ou proativa. Na descoberta reativa, o usuário tem uma idéia vaga do que pode ser a solução ou de onde se pode encontrá-la. Pode-se afirmar que o usuário possui algumas hipóteses iniciais que direcionarão o processo de descoberta. Então, é necessário que haja algum tipo de pré-processamento, no sentido de selecionar atributos ou valores de atributos. Para tanto, exige entender qual o interesse ou objetivo do usuário para limitar o espaço de busca na entrada ou filtrar os resultados na saída (LOH & OLIVEIRA, 2000).

Na descoberta proativa, ao contrário da reativa, a solução do problema é encontrada automaticamente, sem a intervenção do usuário. Segundo Loh & Oliveira (2000), uma expressão comum para definir o modo proativo é: “diga-me o que há de relevante nesse conjunto de dados”.

Nesta dissertação os dois modos de descoberta são usados. Ao avaliar os registros de ocorrência passados e validá-los, aplica-se a descoberta reativa. Ao aplicar as regras geradas pela árvore de decisão, o modo de descoberta é utilizado será o proativo, ou

seja, com as regras, a PMSC pode validar se natureza de operação informada é coerente com a descrição do registro de ocorrência, no momento da inclusão na base de dados.

2.3 Tarefas mais comuns de descoberta de conhecimento em textos

Encontra-se na literatura as seguintes tarefas: classificação ou categorização, sumarização, *clustering* e regras de associação. Qualquer um dos métodos de descoberta tradicional pode ser aplicado nos textos, principalmente se for utilizado o método de extração de informações, que identifica informações relevantes nos documentos e transforma-as em um formato estruturado. Nesta dissertação optou-se pelo método de classificação para auxiliar no processo de mineração dos dados.

2.3.1 Classificação ou categorização

Classificação ou categorização podem ser considerados processos análogos, porém, alguns autores preferem distinguí-los. Esses autores consideram a categorização como sendo um processo que identifica as categorias (nesse caso consideradas como sendo um assunto ou tema) que um documento contém ou se enquadra. Esse processo é basicamente similar ao de classificação, mas sua aplicação é um pouco diferente, já que o primeiro identifica a classe a que o documento pertence e o segundo identifica quais são os assuntos do documento.

Segundo Manning & Schütze (1999), “Classificação ou categorização é a tarefa de atribuir objetos de um universo a duas ou mais classes ou categorias.”

A classificação segundo Yang & Liu (1999):

> É uma técnica empregada para identificar qual classe ou categoria determinado documento pertence, utilizando como base o seu conteúdo. Para tanto, as classes devem ter sido previamente modeladas ou descritas através de suas características, atributos ou fórmula matemática.

É uma das tarefas mais referenciadas na literatura. É também denominada de aprendizado supervisionado, pois a entrada e a saída desejadas são fornecidas previamente por um supervisor externo (FAUSETT, 1994). Por exemplo, pessoas

podem ser previamente grupadas nas classificações de bebês, crianças, adolescentes, adultos e idosos. Dois anos ou menos pode ser mapeado para a categoria bebê.

Podem ser utilizados por qualquer sistema ou técnica que necessite de uma pré-filtragem das informações, tais como sistemas de extração de informações, sistemas de recomendação de informações e leitores de e-mails ou de notícias eletrônicas.

Segundo Brazdil (2004) as principais técnicas de classificação são cinco:

> a) árvore de decisão e regras (ID3, C4.5,C5, CART, CN2,etc);
> b) métodos de regressão adaptados para a classificação (Discriminantes lineares, quadráticos e logísticos);
> c) métodos não-lineares de regressão (Redes neuronais - Back propagation, LVQ etc.);
> d) métodos baseados em instâncias e casos (k-nearest neighbour, kNN);
> e) métodos Probabilísticos (Naïve Bayes, Probabilistic (bayesian) networks).

O objetivo da classificação é construir um modelo que seja capaz de gerar classificações para novos objetos ou novos dados. Para tanto, devem ser considerados dois tipos de atributos:

a) preditivos, cujos valores irão influenciar no processo de determinação da classe;
b) objetivos, que indicam a classe a qual o objeto pertence.

A principal técnica utilizada para a tarefa de classificação é a árvore de classificação ou árvore de decisão. Na Figura 2 pode-se ter a representação visual de uma árvore de decisão utilizada no processo de determinação da natureza da operação policial em função das palavras contidas na descrição da ocorrência.

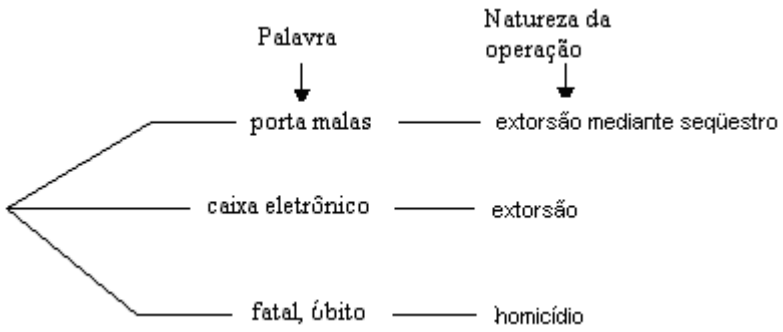


Figura 2 - Exemplo de árvore de decisão

2.4 Árvore de decisão

Segundo Han & Kamber (2001), “uma árvore de decisão é um fluxograma com estrutura tipo árvore, onde cada nó interno denota a realização de um teste em um atributo, cada ramo representa um resultado final do teste, e nós folhas representam classes ou distribuição de classes. O nó mais alto na árvore é a raiz.”

Entre as vantagens na utilização de árvores de decisão, pode-se destacar:

a) pouco tempo de processamento utilizado;
b) facilidade de compreensão do modelo;
c) facilidade em identificar atributos chaves no processo;
d) facilidade de expressar as regras lógicas.

A desvantagem está na instabilidade, ou seja, mudanças no *corpus* de treinamento podem provocar grandes alterações no modelo aprendido pela árvore.

Segundo Kimball (1998 apud Jesus, 2004, p. 39):

> Uma das principais vantagens das árvores de decisão é que o modelo é bem explicativo, uma vez que tem a forma de regras explícitas. Isso permite às pessoas avaliarem os resultados, identificando atributos-chave no processo. As próprias regras podem ser expressas facilmente como declarações lógicas em uma linguagem como Structured Query Language – SQL, de modo que possam ser aplicados diretamente em novos registros.

O processo de geração da árvore de decisão pode ser dividido em duas fases:

a) fase 1: um modelo é construído, descrevendo um conjunto pré-determinado de classes. Em seguida, um *corpus* de treinamento é analisado por um algoritmo de classificação, que gera como saída um modelo baseado numa árvore de decisão, veja Figura 3;
b) fase 2: o modelo gerado pela fase 1 é utilizado para classificação. Depois disso, é realizado um teste e se este for aceitável, as regras poderão ser utilizadas para a classificação de novos casos.

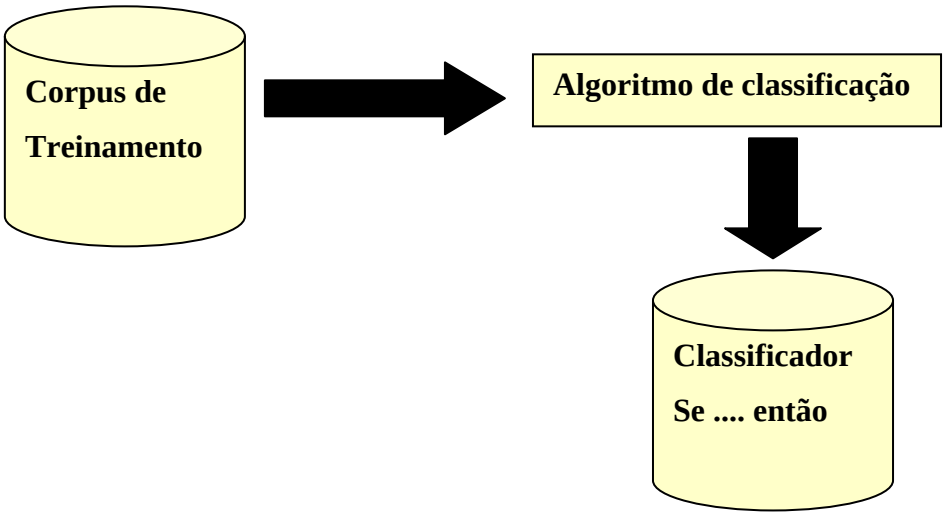


Figura 3 – Construção do modelo utilizando o *corpus* de treinamento

Amplamente utilizadas em algoritmos de classificação, as árvores de decisão são representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados.

As árvores de decisão são formadas de nós e ramos. Os nós representam os atributos e o ramos recebem os valores possíveis desses atributos. Os nós determinados como nós folha representam as categorias (natureza de operação) da base de treinamento. A árvore de decisão divide recursivamente a base de treinamento até que cada subconjunto contenha registros de ocorrência de uma única natureza de operação.

Para atingir este objetivo, a árvore de decisão examina e compara a distribuição das categorias durante a construção da árvore. O resultado obtido, após a construção da árvore de decisão, é utilizado para classificar novos casos.

> Brazdil (2004), afirma que: “Muitos são os algoritmos de classificação que elaboram árvores de decisão. Não há uma forma de determinar qual é o melhor algoritmo, um pode ter melhor desempenho em determinada situação e outro algoritmo pode ser mais eficiente em outros tipos de situações”.

Após a construção de uma árvore de decisão é importante avaliá-la. Esta avaliação é realizada através da utilização de dados que não tenham sido usados no treinamento.

Segundo Manning e Schütze (1999):

> Uma regra é criada para cada caminho desde a raiz até um nódulo da folha. Cada par atributo-valor ao longo de um dado caminho forma uma conjunção no antecedente da regra (a parte do SE). O nó da folha mantém a previsão de classe, formando o conseqüente da regra (a parte do ENTÃO). As regras SE-ENTÃO podem ser mais fáceis para os humanos entenderem, particularmente se a árvore dada for muito grande.

Árvores de decisão podem ser facilmente convertidas para regras de classificação. O conhecimento aprendido pela árvore de decisão pode ser extraído através de regras Se-Então.

Com base na árvore de decisão apresentada na Figura 2 da página 30, pode-se exemplificar a geração de regras. Três exemplos de regras obtidas a partir desta árvore são mostrados a seguir:

a) **Se** palavra = “porta malas” **então** natureza=”extorção mediante seqüestro”;
b) **Se** palavra =”caixa eletrônico” **então** natureza=”extorção”;
c) **Se** palavra=”fatal” ou palavra = “óbito”**então** natureza**=”**homicídio**”.**

2.5 Algoritmo ID3 para construção da árvore de decisão

Foi desenvolvido por Ross Quinlan em 1983, e apresenta resultados práticos. Tem como objetivo formar uma árvore de decisão que classifique uma lista de exemplos, ou seja, a partir de um conjunto de exemplos o algoritmo induz regras do tipo SE...ENTÃO onde cada regra corresponde a um caminho da árvore de decisão.

O conjunto de exemplos é a representação do problema e tem a forma de uma matriz, onde cada coluna é uma característica, ou atributo do problema e cada linha descreve um exemplo através dos valores dos atributos e sua conclusão, ou classificação.

A qualidade das regras produzidas pelo algoritmo ID3 depende diretamente da qualidade do conjunto de dados utilizados para o treinamento da árvore. Os dados utilizados para que a árvore aprenda as regras deve estar totalmente correta ou a árvore não produzirá boas regras. O conjunto de treinamento deve estar completamente pronto e disponível no início do processamento, visto que o ID3 não é um algoritmo incremental. A grande vantagem deste algoritmo é a capacidade de gerar uma árvore de decisão a partir de poucos exemplos. Para selecionar o melhor atributo para ser estudado, é utilizado o critério da entropia.

2.5.1 Critério utilizado para a seleção dos atributos para particionamento

A inovação do algoritmo básico proposto por Quilan em 1986 para o ID3 consiste no critério de seleção dos atributos para o particionamento. Para o ID3 usa-se o critério da entropia. Nesse critério são calculados a entropia e o ganho de informação para cada atributo.

O critério da entropia consiste em medir o grau de aleatoridade dos valores que uma variável X pode assumir. Quanto maior a entropia, maior a aleatoridade dos valores de X. Segundo Han & Kamber (2001, p286), a quantidade de informação esperada para determinar a classificação de uma determinada amostra é determinada por:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Onde:

I(S1, S2, ..., Sm) = é a informação (entropia) necessária para classificar um registro numa das categorias da variável objetivo.

pi = a probabilidade de um registro pertencer a i-ésima classe da variável objetivo, sendo calculado por $p_i = s_i / S$.

s_i = número total de amostras na categoria i.

S = número total de amostras.

Utiliza-se o logaritmo na base 2 já que a informação é utilizada em partes.

Segundo Han & Kamber (2001, p.287) a equação da entropia é determinada como sendo:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj})$$

Onde:

E(A) = é a entropia ou ganho de informação com a partição do atributo A ou devido ao atributo A.

m = representa o número de classes da variável objetivo.

j=1,...,v = representa o número de categorias da variável preditora.

(S1*j*+...+S*mj*) / S = é um peso e corresponde ao numero de amostras na categoria J da variável preditora, dividido pelo número total de amostras.

I(S1*j*,...,S*mj*) = é a informação esperada para a categoria j da variável preditora.

A entropia é máxima quando as classes são equiprováveis, ou seja, existe uma grande aleatoriedade dos dados, neste caso, a entropia calculada é próxima ou maior de 1. A entropia é mínima e igual a 0 quando todos os exemplos pertencem à mesma classe. Quando a entropia é próxima de 0,5 ela é considerada simétrica, conforme o Figura 4.

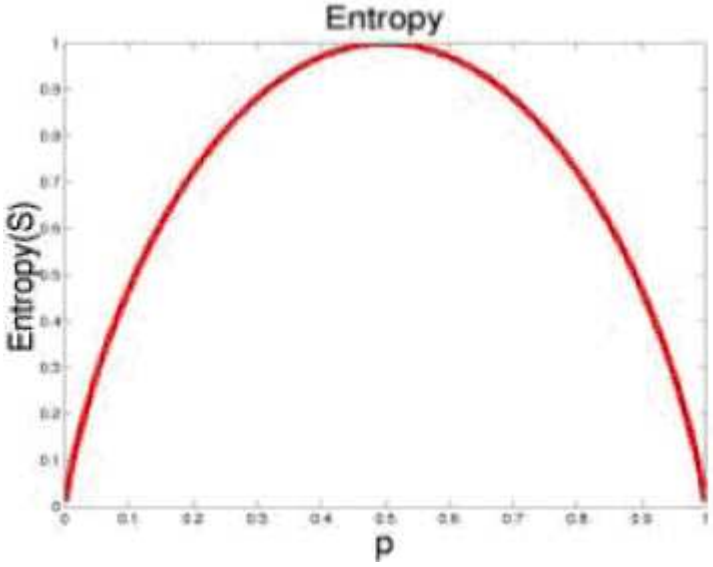


Figura 4 – Representação da entropia

O ganho da informação é a redução esperada da entropia. Deve-se calcular para cada atributo, o seu ganho de informação, ou seja, o quanto ajuda a separar as classes. O ganho da informação é determinanda, segundo Han & Kamber (2001, p.287), pela seguinte equação:

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

No apêndice A, calculou-se a entropia e ganho da informação para a *keyword* Vital. Conforme o anexo C, aonde estão os ROs utilizados como base de treinamento,

calculou-se a entropia de saída (entropia da natureza de operação) e seu ganho de informação (para a *Keyword* vital), conforme demonstra a Figura 5:

Base de testes				
Número de Registros de Ocorrência				121
Natureza	Número de RO por Natureza	Número de RO por Natureza / Número de Registros (p_i)	$\log_2 p_i$	$p_i \log_2 p_i$
A203	1	0,008264463	-6,91886	0,057181
C104	20	0,165289256	-2,59694	0,429245
C112	3	0,024793388	-5,3339	0,132245
C114	5	0,041322314	-4,59694	0,189956
C207	1	0,008264463	-6,91886	0,057181
C215	1	0,008264463	-6,91886	0,057181
C221	13	0,107438017	-3,21842	0,345781
C222	35	0,289256198	-1,78958	0,517647
C610	2	0,016528926	-5,91886	0,097832
D205	4	0,033057851	-4,91886	0,162607
D305	2	0,016528926	-5,91886	0,097832
D309	19	0,157024793	-2,67094	0,419403
D316	2	0,016528926	-5,91886	0,097832
E111	2	0,016528926	-5,91886	0,097832
E112	10	0,082644628	-3,59694	0,297267
E123	1	0,008264463	-6,91886	0,057181
	121			3,114205

	Vitais			Para vital=1			Para vital=0		
	sim(1)	não(0)	total	pi1	log2(pi1)	pi1*log2(pi1)	pi0	log2(pi0)	pi0*log2(pi0)
A203	0	1	1	0	0	0	0,01098901	-6,50779	-0,0715142
c104	5	15	20	0,166666667	-2,584962501	-0,4308271	0,16483516	-2,6009	-0,4287204
c112	0	3	3	0	0	0	0,03296703	-4,92283	-0,1622912
c114	1	4	5	0,033333333	-4,906890596	-0,163563	0,04395604	-4,50779	-0,1981448
c207	0	1	1	0	0	0	0,01098901	-6,50779	-0,0715142
c215	0	1	1	0	0	0	0,01098901	-6,50779	-0,0715142
c221	0	13	13	0	0	0	0,14285714	-2,80735	-0,4010507
c222	0	35	35	0	0	0	0,38461538	-1,37851	-0,5301968
c610	1	1	2	0,033333333	-4,906890596	-0,163563	0,01098901	-6,50779	-0,0715142
d205	1	3	4	0,033333333	-4,906890596	-0,163563	0,03296703	-4,92283	-0,1622912
d305	2	0	2	0,066666667	-3,906890596	-0,2604594	0	0	0
d309	19	0	19	0,633333333	-0,658963082	-0,4173433	0	0	0
d316	0	2	2	0	0	0	0,02197802	-5,50779	-0,1210504
e111	0	2	2	0	0	0	0,02197802	-5,50779	-0,1210504
e112	1	9	10	0,033333333	-4,906890596	-0,163563	0,0989011	-3,33787	-0,330119
e123	0	1	1	0	0	0	0,01098901	-6,50779	-0,0715142
	30	91	121			-1,7628818			-2,8124861
						1,76288182			2,81248606
	E(Vital)=		2,552253						
	Ganho(Vital)=		0,561952						

Figura 5 – Cálculo da entropia e ganho de informação da base de treinamento

2.5.2 Pseudocódigo do algoritmo ID3

Segundo Brasil (2004, p. 47) o algoritmo ID3 usa atributos booleanos, que gera apenas duas ramificações, sendo o pseudocódigo apresentado a seguir:

Entrada: um conjunto de exemplos de aprendizado E.
if todos os exemplos de E satisfazem a condição de término t(E) then
    Retorne o valor da classe
else
    para cada atributo a_i determine o valor da função aval(E, a_i). Seja a_j o atributo com o melhor valor de aval(E, a_i)
    aj assume o valor 0 ou 1, crie o seguinte nó da árvore
    retire o melhor atributo da lista de possíveis atributos e particione os exemplos do conjunto E nos subconjuntos E_0 (correspondente aos valores de $a_j = 0$) e E_1 (correspondente aos valores de $a_j = 1$)
    aplique o algoritmo recursivamente para os subconjuntos E_0 e E_1
end if

2.5.3 Poda da árvore de decisão

Após construir a árvore de decisão, é possível que ela esteja muito específica para o conjunto de treino utilizado e não classifique bem os objetos do conjunto de teste diz-se que a árvore “decorou” os dados de treino (*overfitting*).

Para evitar a situação descrita, os algoritmos indutores acrescentam uma fase de poda da árvore construída, ou seja, a remoção de alguns nós da árvore aumentando a generalização da mesma. A execução da poda, pode ser realizada de 2 formas:

a) pré-poda: executado durante o processo de indução da árvore, impõe um *threshold* para a proporção da classe mais freqüente para o qual um nó é forçado ser folha, não prosseguindo o particionamento por este ramo. Exemplo: se a classe mais freqüente responde por mais de 70% dos objetos em uma partição, este nó não será mais particionado;

b) pós-poda: executada após a árvore ter sido construída, a partir dos nós-folha, a sub-árvore formada por eles e seu nó-pai é analisada. Se a taxa de erros de classificação for reduzida perante uma substituição da sub-árvore por um único nó terminal, então a árvore é podada nesta parte.

2.6 *Software Weka*® para a construção da árvore de decisão

Segundo *Weka* (2005), o pacote *Weka*® (*Waikato Environment for Knowledge Analysis*) é formado por um conjunto de implementações de algoritmos de diversas técnicas de mineração de dados. Desenvolvido na linguagem Java, que tem como principal característica a portabilidade, esse algoritmo pode rodar nas mais variadas plataformas (*Windows, Linux, Mac Os,* etc). Além disso, é um *software* de domínio público, disponível em http://www.cs.waikato.ac.nz/ml/weka/. Ao se instalar o *Weka*®, será apresentada uma tela inicial com quatro botões, Figura 6.

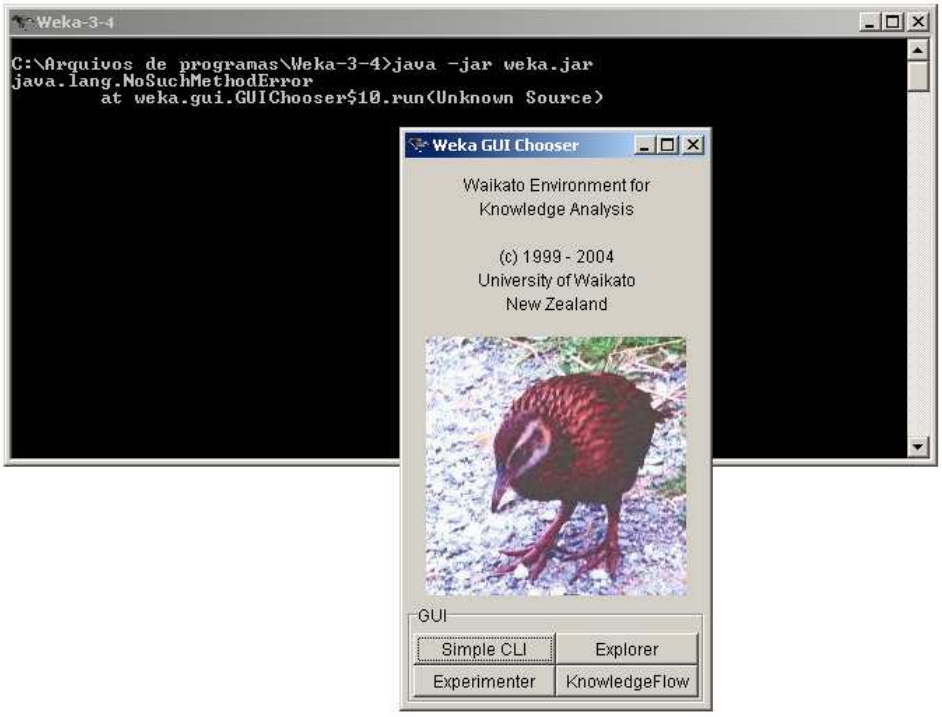


Figura 6 – Tela inicial do pacote *Weka*®

O *Weka*® implementa vários métodos de classificação: Árvore de Decisão, Regras de Aprendizagem, Naive Bayes, Tabelas de Decisão, Regressão Lógica, SVM

(*Support Vector Machines*), entre outros. Os algoritmos disponíveis para geração de árvore de decisão, podem ser visualizados na Figura 7:

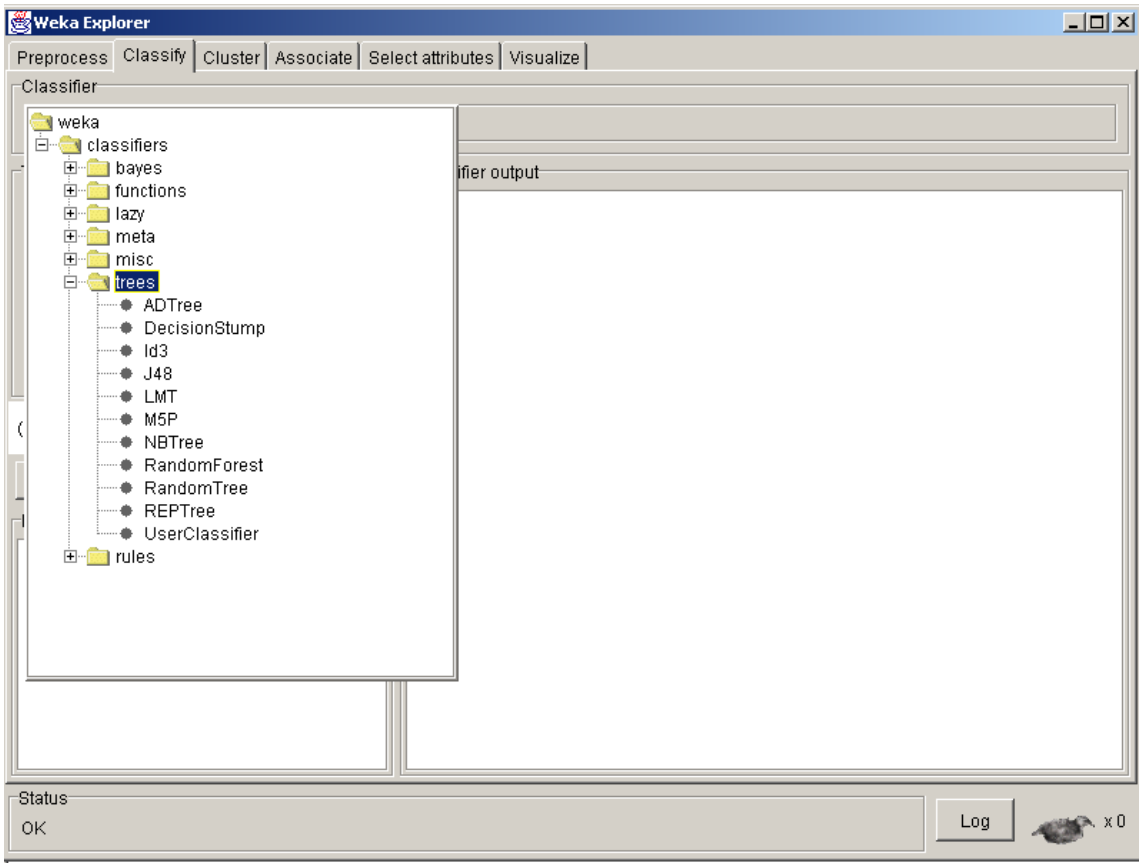


Figura 7 – Algoritmos implementados pelo *Weka*®

O *Weka*® possui um formato próprio para o arquivo de entrada de dados, o ARFF. Antes de aplicar os dados a qualquer algoritmo do pacote *Weka*® estes devem ser convertidos para o formato ARFF.

2.6.1 Arquivo ARFF

O formato ARFF consiste basicamente de duas partes. A primeira contém uma lista de todos os atributos (Real, *Integer*, *String*, etc), definidos por um tipo ou por um conjunto de valores. Se utilizarmos, os valores estes devem estar entre "{ }" separados por vírgula. A segunda parte consiste das instâncias, ou seja, os registros a serem minerados com o valor dos atributos para cada instância separado por vírgula, a

ausência de um item em um registro deve ser atribuída pelo símbolo de interrogação "?". Na primeira linha, deve conter o comando @relation (nome do conjunto de dados), em seguida tem-se a relação dos atributos, onde coloca-se o nome do atributo e tipo ou seus possíveis valores, definido por @attribute nome_do_atributo tipo ou {valores}. Após isso deve conter o comando @data e nas próximas linhas devem ser listados os registros onde cada linha representa um registro. O exemplo, na Figura 8, é instalado junto com o pacote *Weka*®. O arquivo contém dados atmosféricos tais como: temperatura, umidade, vento e como estava o céu. Esses dados foram coletados durante 14 partidas de golfe, e servirão para gerar regras, definindo se deve ou não jogar golfe. A variável objetivo é joga, podendo apresentar os seguintes resultados: sim ou não.

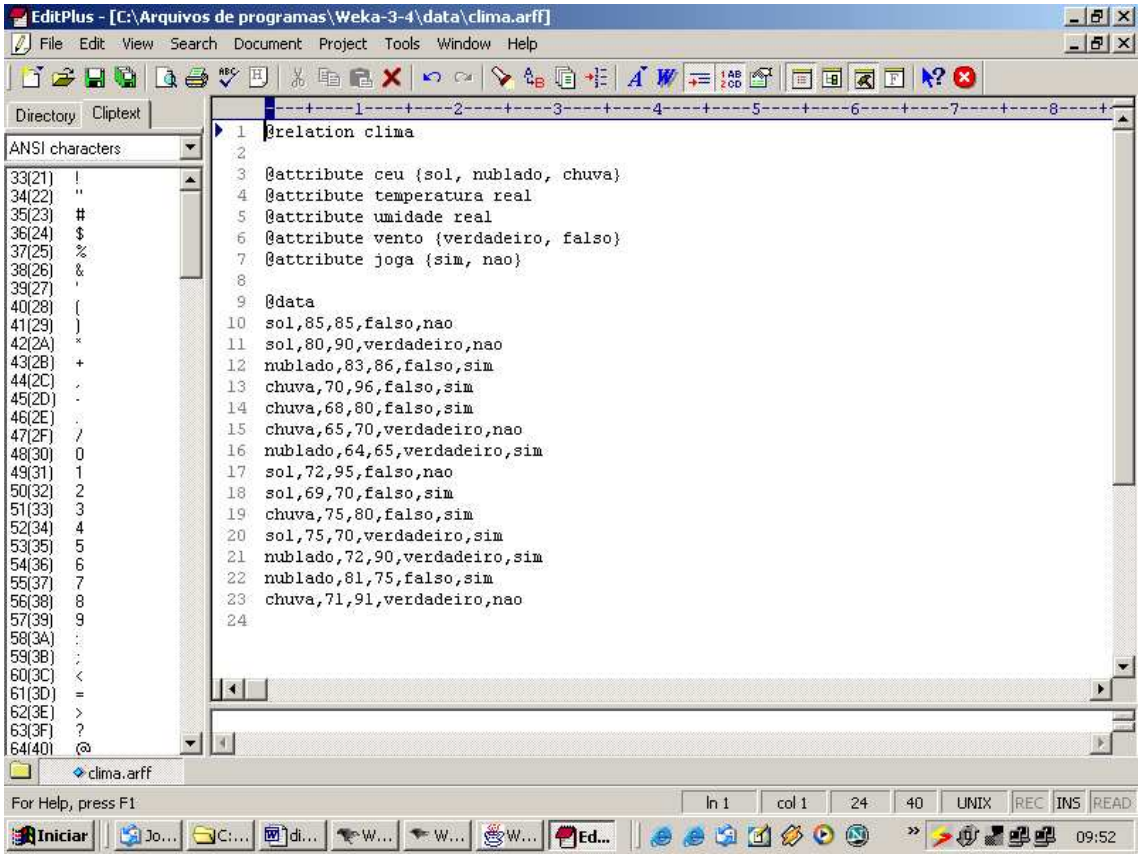


Figura 8 – Arquivo no formato ARFF

Para iniciar o *Weka*® pressiona-se o botão *Explorer*, a janela *Weka Knowledge Explorer* será aberta, deve-se então carregar os dados para serem analisados os quais podem ser originados de um arquivo (*Open file*...) de uma URL (*Open URL*...) ou ainda de um banco de dados (*Open DB*...). Neste exemplo os dados encontram-se em um

arquivo ARFF, clicando em *Open File* carrega-se o arquivo de dados, conforme ilustra a Figura 9.

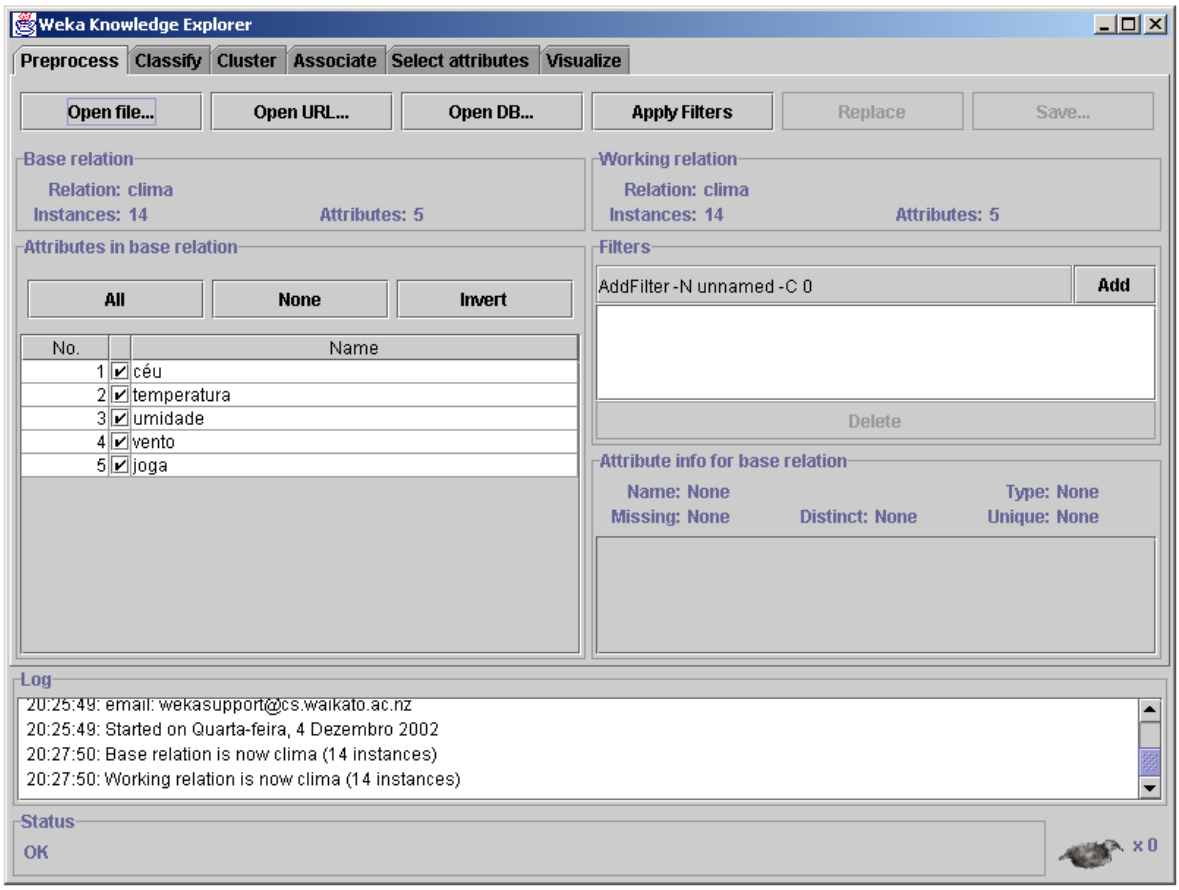


Figura 9 – Carregando o arquivo ARFF

Na parte superior encontra-se as seguintes abas: *Preprocess*, onde se pode editar e salvar a base; *Classify*, conjunto de algoritmos que implementam os esquemas de aprendizagem que funcionam como classificadores; *Cluster*, contém os algoritmos para geração de grupos; *Associate*, conjunto de algoritmos para gerar regras de associação, *Select attributes* determina a relevância dos atributos e *Visualise* explora os dados.

Em *Test options* definem-se algumas opções de teste como conjunto de treinamento (*Use training set*), fornecer um conjunto de teste (*Supplied test set*), validação cruzada (*Cross-validation*) com o número de partições e porcentagem dos dados usados para treinamento (*Percentage split*) em *More options*... temos algumas opções de saída.

No exemplo, como se tem poucos registros, utilizaram-se os dados como um conjunto de treinamento ativando a opção *Use training set*. Selecionou-se o algoritmo J48[4] e clicou-se em *Start*, o *Weka*® gerou o seguinte resultado:

```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     clima
Instances:    14
Attributes:   5
              ceu
              temperatura
              umidade
              vento
              joga
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===
J48 pruned tree
------------------
ceu = sol
|   umidade <= 75: sim (2.0)
|   umidade > 75: nao (3.0)
ceu = nublado: sim (4.0)
ceu = chuva
|   vento = verdadeiro: nao (2.0)
|   vento = falso: sim (3.0)
Number of Leaves  :   5
Size of the tree :    8

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          9               64.2857 %   (9 / 14)
Incorrectly Classified Instances        5               35.7143 %  (5 / 14)
Kappa statistic                         0.186

=== Detailed Accuracy By Class ===
```

[4] O algoritmo J48 é uma implementação do algoritmo C4.5, ambos utilizados para a mineração de documentos textuais. Ele constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, e usa esse modelo para classificar outras instâncias num conjunto de testes.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.778	0.6	0.7	0.778	0.737	sim
0.4	0.222	0.5	0.4	0.444	não

=== Confusion Matrix ===

a	b	<-- classified as
7	2	a = sim
3	2	b = não

Na Figura 10 ilustra-se a árvore de decisão seguindo as regras geradas pelo *Weka*®.

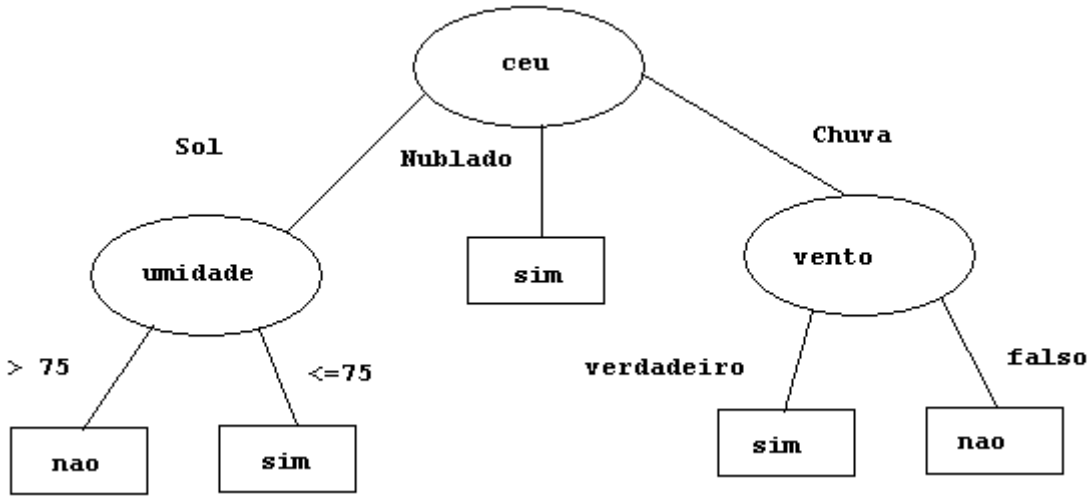


Figura 10 – Árvore de decisão gerada pelo programa *Weka*®

Com base na árvore de decisão apresentada na Figura 10, pode-se extrair as seguintes regras:

a) Se ceu = sol e umidade > 75 então jogar = não;
b) Se ceu = sol e umidade <= 75 então jogar = sim;
c) Se ceu = nublado então jogar = sim;
d) Se ceu = chuva e vento = verdadeiro então jogar = sim;
e) Se ceu = chuva e vento = falso então jogar = não.

Conclui-se que em 2 situações não será permitido jogar:

a) toda vez que estiver fazendo sol e a umidade relativa do ar for menor que 75%;
b) se estiver chovendo e não tiver vento.

Várias interpretações podem ser obtidas pelos dados apresentados como resultado, por exemplo, a taxa de erro no caso de permissão para jogar consiste em 77,78% e para não jogar é de 40%.

2.6.2 Resultados obtidos pelo *Weka*®

A avaliação de um modelo é realizada sobre o conjunto de teste. Em aprendizado de máquina, os métodos usados para seleção dos exemplos são *hold-out* e *cross-validation*, sendo que ambos estimam a qualidade média do modelo construído por meio de várias amostragens. A seguir algumas definições:

Hold-out: um conjunto de teste de tamanho pré-definido é selecionado aleatoriamente, sendo que o restante dos exemplos do conjunto é usado para treinamento do modelo.

Cross-validation: o conjunto de dados é particionado em N subconjuntos sendo que as fases de teste são realizadas em N vezes. Em cada interação, um desses subconjuntos é selecionado para ser o conjunto de teste, enquanto os demais compõem o conjunto de treinamento.

Na classificação que abrange classes de valores discretos, as estimativas de qualidade podem ser determinadas por meio de uma matriz de confusão obtida a partir dos resultados da classificação. A matriz de confusão envolve os valores reais dos exemplos e os valores preditos pelo modelo. Na Tabela 1 é apresentado um modelo de classificação bi-valorada, na qual tp representa a quantidade de exemplos da classe positiva que foram classificados como positivos (positivo verdadeiro), fp representa a quantidade de exemplos da classe negativa que foram classificados como positivos (falso positivo), fn representa a quantidade de exemplos da classe positiva que foram classificados como negativos (falso negativo), e tn representa a quantidade de exemplos da classe negativa que foram classificados como negativos (negativo verdadeiro).

Tabela 1 - Modelo de Classificação Bi-valorada

	Predito como positivo	**Predito como negativo**
Exemplos positivos	*tp*	*fp*
Exemplos negativos	*fn*	*tn*

Partindo-se do exemplo do clima para jogar golfe, tem-se a seguinte matriz de referência cruzada:

Predição	Sim	Não	
Exemplos Positivos	7	3	10
Exemplos negativos	2	2	4
Total	9	5	14

A partir da matriz de referência cruzada, pode-se obter vários resultados para medir a qualidade da classificação dos textos. Esses resultados são obtidos das seguintes medidas de avaliação: *Precision, Recall, F-measure* e *Kappa*, que são definidos abaixo:

Precision: a precisão de um modelo é a proporção de exemplos positivos que foram corretamente classificados. Pode ser calculada de acordo com a equação abaixo:

$$Precision = tp / (tp + fp)$$

Onde:

tp = positivo verdadeiro

fp = falso positivo

Recall: é definida como a porção classificada corretamente como exemplos positivos. A estimativa dessa medida é calculada conforme a seguinte a equação:

$$Recall = tp / (tp + fn)$$

Onde:

tp = positivo verdadeiro

fn = falso negativo

F-measure: essa medida, também conhecida por medida F, combina *Precision* e *Recall*. Essa medida é calculada pela seguinte fórmula:

$$F\text{-}measure = \frac{2 \times Pr \times Rc}{Pr + Rc}$$

Onde:

Pr = *Precision*

Rc = *Recall*

Kappa: é uma medida de concordância entre os dados. Esta medida de concordância tem como valor máximo 1, onde este valor 1 representa total concordância e os valores próximos e até abaixo de 0, indicam nenhuma concordância. Agresti (apud LANDIS & KOCH, 1977) sugerem a Tabela 2 para a interpretação dos valores de *Kappa*.

Tabela 2 - Interpretação dos valores de *Kappa*

Valores de *Kappa*	Interpretação
<0	Nenhuma concordância
0-0.19	Baixa concordância
0.20-0.39	Concordância média/baixa
0.40-0.59	Concordância moderada
0.60-0.79	Condordância significativa
0.80-1.00	Concordância quase perfeita

Essa medida é calculada pela seguinte fórmula:

$$kappa = \frac{po - pe}{1 - pe}$$

po=probabilidade de concordância

pe=é a probabilidade esperada se as proporções forem independentes.

Aplicando-se as medidas de avaliação descritas acima, tem-se:

po = (7/14 +2/14) = 0,642857

pe = (9/14 * 10/14 + 5/14 * 4/14) = 0,561225

tp = 7

tf = 3

fn = 2

Precision = 7 / (7 + 3) = 0,70

Recall = 7 / (7 + 2) = 0,778

F-measure = 2 * 0,70 * 0,778 / (0,778 + 0,70) = 0,737

$$Kappa = \frac{0,642857 - 0,561225}{1 - 0,561225}$$

$$Kappa = \frac{0,081632}{0,438775}$$

Kappa = 0,186

Conforme a Tabela 2, o valor 0,186 indica baixo grau de concordância.

3 MÉTODOS E TÉCNICAS DE PESQUISA

O método de pesquisa é o conjunto de atividades sistemáticas e racionais, que, com maior segurança permite atingir os objetivos propostos, indicando o caminho a ser seguido (LAKATOS & MARCONI, 2003). Esta investigação utiliza a pesquisa bibliográfica e a pesquisa descritiva, exploratória e qualitativa. Na primeira fase, é construída uma pesquisa bibliográfica, que, segundo Gil (1999), é desenvolvida a partir de material já elaborado, constituído de livros e artigos científicos.

Neste estudo, foram utilizados materiais existentes sobre *text mining*, como livros, periódicos, tese, dissertações e documentos extraídos de sites especializados no assunto, que estão relacionados nas referências bibliográficas dessa dissertação.

Mattar (1999) menciona que para se desenvolver uma pesquisa descritiva torna-se necessário um profundo conhecimento do tema a ser estudado, necessitando que o pesquisador saiba o que deseja, ou seja, o que pretende validar, quando, onde, como e quem o fará.

Como a pesquisa ocorre em população previamente definida (os registros de ocorrência policiais da Região Metropolitana da Grande Florianópolis no ano de 2003) pretende-se desenvolver uma metodologia com base em uma descrição textual dos registros de ocorrência e validar o modelo proposto. A pesquisa descritiva é escolhida, pois, como assegura Gil (1999), essa tem o objetivo de estudar as características de um determinado grupo. Outro fator desta pesquisa que vem ao encontro da pesquisa descritiva é a utilização de técnicas padronizadas de coleta de dados (ROESCH, 1999). Esta pesquisa utiliza arquivos em forma de textos para a obtenção dos dados primários.

Gil (1999) relata que as pesquisas exploratórias são desenvolvidas com o objetivo de proporcionar uma visão geral sobre o objeto da pesquisa, este instrumento escolhido quando o assunto é pouco explorado. Segundo Mattar (1999) o levantamento bibliográfico é um método utilizado na pesquisa exploratória.

3.1 Segurança Pública

Para um melhor entendimento sobre a constituição da segurança pública, é interessante compreender alguns termos que estão relatados no site http://www.pm.sc.gov.br da Secretaria de Segurança Pública do Governo Federal.

Segue alguns conceitos:

> A DEFESA SOCIAL inclui, entre outras atividades, a prestação de serviços de segurança pública e de defesa civil.
>
> A SEGURANÇA PÚBLICA é uma atividade pertinente aos órgãos estatais e à comunidade como um todo, realizada com o fito de proteger a cidadania, prevenindo e controlando manifestações da criminalidade e da violência, efetivas ou potenciais, garantindo o exercício pleno da cidadania nos limites da lei.
>
> A DEFESA CIVIL é um conjunto de medidas que visam prevenir e limitar, em qualquer situação, os riscos e perdas a que estão sujeitos a população, os recursos da nação e os bens materiais de toda espécie, tanto por agressão externa quanto em conseqüência de calamidades e desastres da natureza.
>
> As POLÍCIAS MILITARES são os órgãos do sistema de segurança pública aos quais competem as atividades de polícia ostensiva e preservação da ordem pública.
>
> As POLÍCIAS CIVIS são os órgãos do sistema de segurança pública aos quais competem, ressalvada competência específica da União, as atividades de polícia judiciária e de apuração das infrações penais, exceto as de natureza militar (BRASIL, 2005).

Em 1813 foi criada a Polícia Militar de Santa Catarina (PMSC), na cidade de Desterro, hoje Florianópolis. Chamava-se de Força Policial. No início, sua principal atividade era defender a costa catarinense, principalmente a Ilha de Santa Catarina.

Em 1860, existem relatos que havia um quartel para a força pública, com sede numa das salas térreas do Palácio do Governo, na praça principal do povoado.

Durante todo esse tempo a PMSC atravessou várias fases. Hoje, 168 anos depois, a PMSC conta com um efetivo de aproximadamente 13.000 (treze mil) homens (SANTA CATARINA, 2004).

A PMSC executa várias tarefas, dentre as quais:

a) o serviço Emergência 190: aqui são atendidos e cadastrados os ROs (objeto desse estudo);
b) policiamento ostensivo a pé: composto por policiais fardados, equipados com ou sem viatura que faz policiamento nas ruas;

c) policiamento motorizado de motocicleta: policiais que trabalham interligados ao Centro de Operações através de rádio comunicação;
d) policiamento ostensivo de trânsito: executa serviços como orientação do tráfego, atendimento e socorro em acidentes, remoção, retenção e apreensão de veículos em situação irregular, fiscalização de documentos de porte obrigatório, autuação por infração de trânsito e participação em campanhas educativas;
e) policiamento com cães e montado;
f) batalhão de operações especiais (BOE);
g) companhia de operações especiais (COE);
h) patrulhamento aéreo;
i) policiamento de proteção ambiental;
j) policiamento rodoviário;
k) policiamento em praias;
l) segurança a dignitários.

Todos esses serviços que a PMSC executa resulta em um registro de ocorrência que será descrito no próximo item.

O aumento da população gerou um indício maior de criminalidade. Os órgãos de Segurança Pública passaram a desenvolver formas de investigação, visando combater a criminalidade.

Em 03/12/2002 foi criada a Diretoria de Combate ao Crime Organizado (DIRC), sendo criada por força de convênio firmado entre o Estado de Santa Catarina e Governo Federal através do Ministério da Justiça, Secretaria Nacional de Segurança Pública. Tem como um de seus objetivos a Integração ao Subsistema de Inteligência de Segurança Pública. O Núcleo de Gerenciamento do Subsistema de Inteligência e Estatística de Segurança Pública como foi intitulado primeiramente, foi pioneiro em sua criação, pois trabalhariam juntos policiais civis e militares. Sua estrutura contaria com um Diretor de Combate ao Crime Organizado, um Gerente de Inteligência e um Gerente de Estatística. (SSPSC, 2004).

A DIRC exerce o papel de identificar e entender os crimes mais complexos assessorando as autoridades governamentais na elaboração de planos e políticas de

Segurança Pública. A Gerência de Inteligência tem como missão produzir conhecimento para a tomada de decisão pró-ativa ou reativa conforme os dados revelem.

As estatísticas criminais, utilizadas pela DIRC, são quase sempre calculadas tendo por base os ROs. Por isso, há uma grande preocupação no preenchimento correto.

Por exemplo: os dados como autor e vítima são imprescindíveis para efeitos judiciais. Outro exemplo da necessidade dos dados serem extremamente confiáveis está no campo de natureza de operação, em que gera vários outros dados estatísticos. Se os ROs estiverem incorretos irão gerar estatísticas falsas que levarão os Diretores a tomarem decisões erradas. Por isso os ROs devem ser fonte confiável para estatísticas criminais

Muitos dessas ocorrências são registradas por telefone, como agressões, depredações, maus tratos, etc, pelas vítimas ou por testemunhas do ato. As centrais de atendimento da PMSC são informatizadas, catalogando um grande número de delitos e trazendo informações adicionais a respeito do tratamento efetuado ao problema.

3.1.1 O que é o registro de ocorrência

O RO é um documento legal elaborado pela Polícia que representa a primeira notificação oficial de uma queixa-crime para a maior parte dos casos que são encaminhados a uma unidade de polícia judiciária. Considerando as especificidades das funções de polícia investigativa e de polícia judiciária, pode-se dizer que o RO expressa o atendimento preliminar oferecido ao público e agentes institucionais que, por diversas motivações, acionam os serviços policiais civis. Trata-se, portanto, de um instrumento artesanal e, em certa medida, versátil, no qual são registrados não só aqueles fatos interpretados juridicamente como crimes e contravenções, como também os atos administrativos efetuados por uma unidade policial distrital e/ou especializada. Por conta disso, o RO consiste na principal ferramenta que aciona boa parte das rotinas executivas, investigativas e cartorárias desenvolvidas em uma delegacia. Dentro da linha de produção do trabalho policial civil (que, de uma forma simplificada, começa no balcão de atendimento, passa pela confecção de depoimentos, levantamento de provas, averiguações, diligências solicitadas, incursões nas ruas e termina com o envio de inquéritos ao Ministério Público).

Segundo (Muniz, 2000, p. 124):

> O registro de ocorrência destaca-se como uma forma de comunicação legal que procura atender, minimamente, a propósitos operacionais diferentes, porém complementares:
> a) orientar o trabalho da Polícia Investigativa, através da coleta de informações iniciais que contribua tanto para a elucidação futura do delito notificado, quanto para a constituição de uma memória investigativa;
> b) subsidiar o trabalho da Polícia Judiciária no que se refere aos procedimentos cartorários previstos no Código de Processo Penal;
> c) registrar bens apreendidos e outros procedimentos administrativos internos e
> d) solicitar e registrar o empenho dos serviços periciais da Polícia Técnica-científica.

3.1.2 Como funciona

Em caso de emergência, o cidadão liga gratuitamente para o telefone 190 (Figura 11), no centro de operações em Florianópolis. A ligação chega à sala de atendimento, onde operadores treinados verificam a origem da chamada, registram a ocorrência e repassam para a sala de despachos. Os despachantes atendem a ocorrência e se comunicam via rádio acionando a unidade. Conforme o caso, as unidades mais próximas e adequadas de rádio patrulha, policiamento ambiental, de trânsito, montada, a pé, rodoviária, etc podem ser deslocadas rapidamente até o local da ocorrência. Os cidadãos também podem registrar as ocorrências pela internet no endereço: http://sistemas.sc.gov.br/bocidadao/ conforme ilustra a Figura 12. Os seguintes registros podem ser efetuados pela internet: perda de documentos, perda de objetos, furto de celular, denúncia anônima, denúncia e ameaça, (Figura 13).

No anexo B constam as 6 etapas para o preenchimento de um registro de ocorrência pela internet.

Figura 11 – Emergência 190 - Centro de Operações da PMSC
Fonte: Santa Catarina (2004)

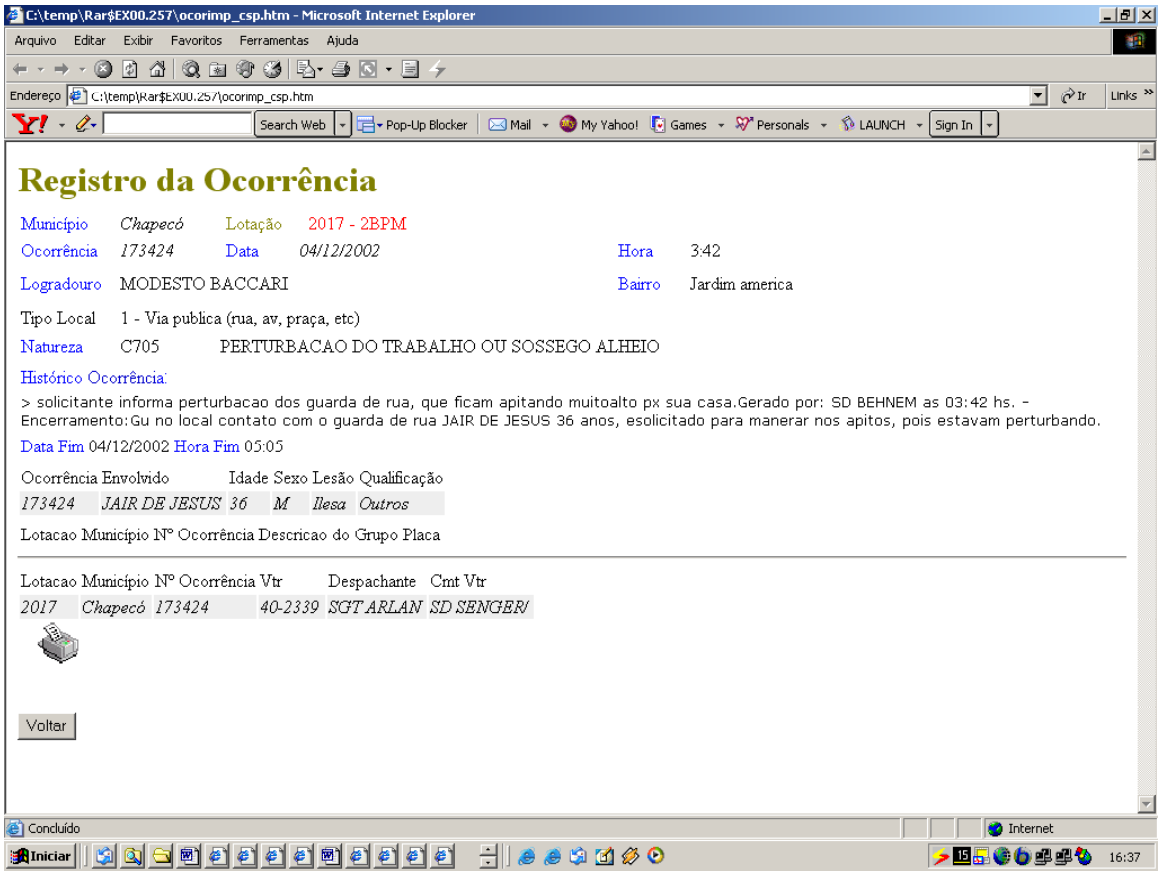


Figura 12 – Registro de ocorrência preenchido via internet

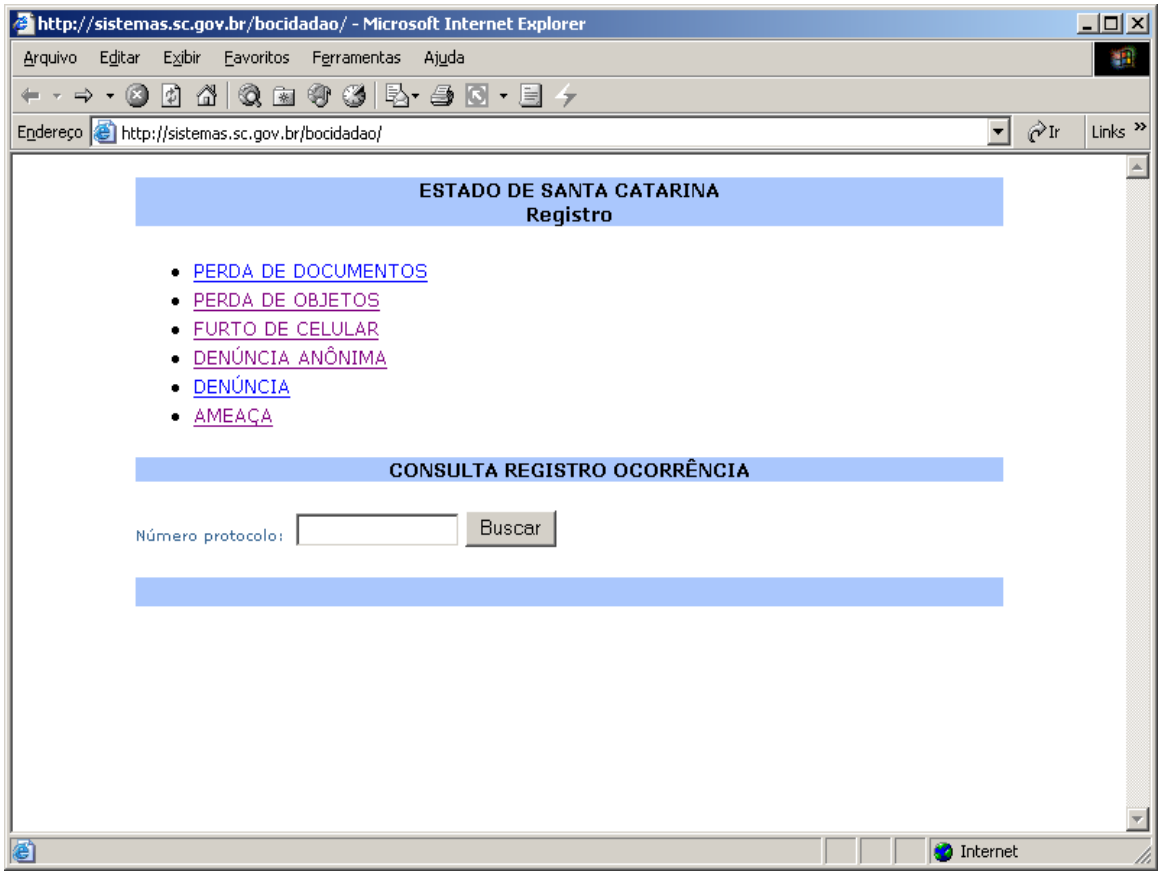


Figura 13 – Página principal do registro de ocorrência pela internet
Fonte: Santa Catarina (2004)

A Figura 14 demonstra o fluxo que o RO dispara ao relatado. Em branco estão relacionados os agentes: sociedade, vítima, agressor, policiais militar e civil, ministério público, Justiça e Sistema Penitenciário. Ao informar um RO boa parte desses órgãos e pessoas são acionados. Cabe ressaltar que nem todos os ROs chegam a abertura de inquérito. Em preto estão relacionados os estágios que o RO passa.

Para exemplificar, imagine que um indivíduo utiliza uma arma para assaltar uma pessoa. Essa pessoa liga para o 190 relatando o fato. A Polícia Militar é acionada para socorrer a vítima e registra a ocorrência. Nesse exemplo o indivíduo é o agressor. A pessoa é a vítima. O instrumento de agressão é a arma. A vítima ao ligar para o 190 faz a denúncia que desencadeia o processo de relato e abertura de inquérito, denúncia, julgamento e penalização do agressor. Como conseqüência a Polícia Civil e Militar necessitam de medidas preventivas. É nesse passo que esse estudo se encaixa, procurando validar de forma eficiente os dados nos ROs para que as medidas preventivas sejam eficazes.

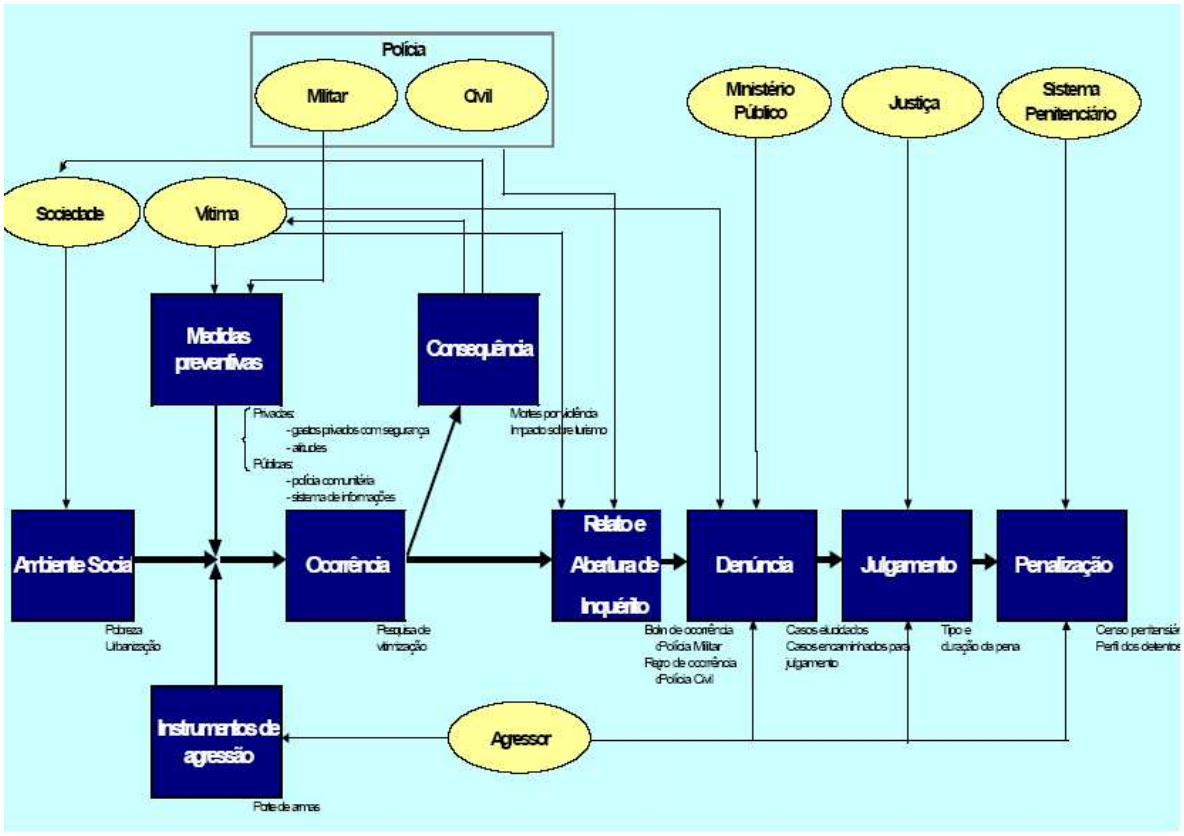


Figura 14 – Fluxo das Informações no Sistema de Segurança e Justiça.
Fonte: Cerqueira et al., (2000).

3.2 Dados fornecidos pela Secretaria de Segurança Pública e Defesa do Cidadão

O universo deste estudo foi constituído dos registros de ocorrências policiais da Grande Florianópolis[5] no ano de 2003, dando-se ênfase nos delitos de apoio/reforço a polícia civil, homicídio, seqüestro e/ou cárcere privado, tentativa de homicídio, extorsão mediante seqüestro, roubo consumado, roubo ou assalto contra pessoa, roubo ou assalto a estabelecimento, disparo de arma de fogo, encontro de cadáver, endereço incompleto, óbito no local, vítima já conduzida por populares, ferimento por arma branca, ferimento por arma de fogo e traumatismo crâneo-encefálico. A base de dados contém 2684 RO.

A pesquisa ficou restrita a este universo, devido as mudanças ocorridas na troca de secretário na Secretaria de Segurança Pública.

[5] Constituída pelos municípios de Biguaçú, Florianópolis, Palhoça, Santo Amaro da Imperatriz e São José.

3.2.1 Variáveis

As variáveis existentes nos registros de ocorrência são: município de ocorrência, lotação, número da ocorrência, data, hora, logradouro, tipo de local, natureza de operação, histórico e/ou descrição da ocorrência, data e hora de encerramento da descrição, envolvido, idade, sexo, lesão, qualificação, descrição do grupo, placa, vtr (viatura), despachante e cmt (comandante).

Vale ressaltar que as variáveis utilizadas neste estudo foram apenas: a natureza de operação, o histórico da ocorrência, a cidade e o número da ocorrência.

3.2.2 Base de dados

Durante o ano de 2004 a DIRC, enviou alguns arquivos nos formatos doc, xls e txt para unificação desses dados a fim de minerá-los e classificá-los corretamente. Elaborou-se um formato padrão para mineração que deve conter: natureza de operação, número da ocorrência e a descrição, conforme demonstra a Figura 22 da página 68. Os registros de ocorrências recebidos estão classificados conforme a Tabela 3.

3.3 Palavras-chave por natureza de operação

A Secretaria de Segurança Pública e Defesa do Cidadão do Estado de Santa Catarina ao enviar os dados para a análise, enviou também quais seriam as palavras-chaves para cada natureza de operação, que podem ser vistas na Tabela 4.

Algumas naturezas de operação não continham palavras-chaves, o que resultou num processo de identificação de palavras-chaves, que será descrito no capítulo 4.

Contabilizou-se 17354 palavras não repetidas.

Tabela 3 - Quantidade de RO por natureza de operação

Natureza de Operação[6]	**Quantidade de registros recebidos**
A203	1
C104	24
C112	9
C114	5
C207	4
C215	2
C221	36
C222	2557
C610	2
C903	1
D205	4
D305	2
D309	21
D316	3
E111	2
E112	10
E123	1
Total de RO	2684

Tabela 4 - Palavras-chaves para cada natureza de operação

Natureza de Operação	**Palavras-chaves**
C207 – Extorsão mediante seqüestro	Porta malas, caixa eletrônico
C104 – Homicídio	Fatal, óbito, sinais vitais
C215 – Roubo consumado	Arma
C221 – Roubo ou assalto contra pessoa	Tomado assalto
C112 – Seqüestro ou cárcere privado	Manteve, relâmpago, seqüestro

[6] Para conhecer o significado de cada código de natureza de operação consulte o anexo A.

3.4 Modelo utilizado para aplicação de *Text Mining* em Segurança Pública

Para o processo de mineração dos dados textuais sobre os dados da Secretaria de Segurança Pública e Defesa do Cidadão será utilizada a metodologia CRISP-DM (2005). A Figura 15 apresenta o modelo proposto por essa metodologia.

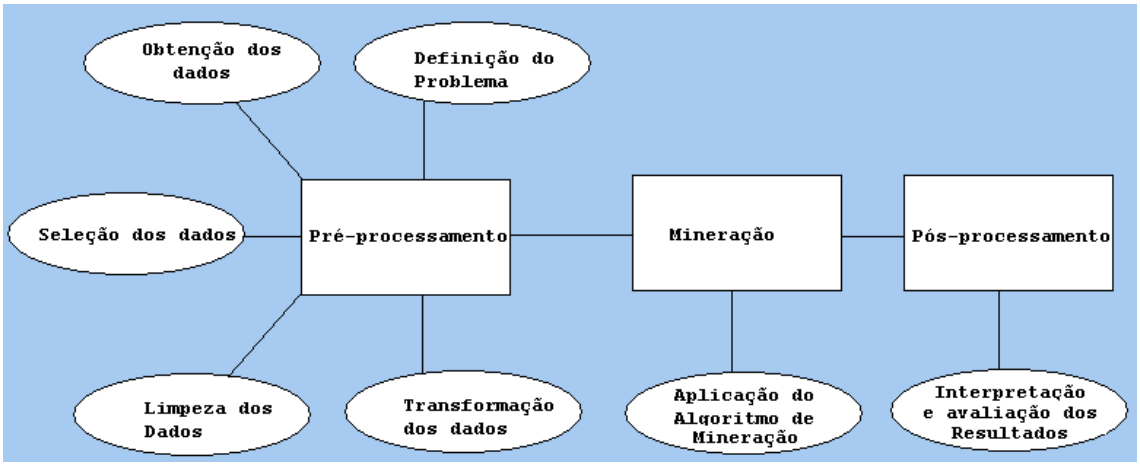


Figura 15 - Modelo do processo de mineração em base de dados textuais para Segurança Pública, adaptado de CRISP-DM, 2005.

A seguir são descritas as etapas do modelo.

3.4.1 Pré-processamento

O entendimento e a definição do problema envolvem a compreensão do domínio e a definição clara dos objetivos da pesquisa, o que proporciona um maior direcionamento do trabalho desenvolvido. O resultado obtido sem a definição e estudo do problema pode ser considerado pouco confiável, mesmo que o desenvolvimento do processo envolva a aplicação de técnicas sofisticadas para a extração de conhecimento.

Essa etapa é composta de cinco tarefas:

a) definição do problema;
b) obtenção dos dados;
c) seleção dos dados;

d) limpeza dos dados;
e) transformação dos dados.

3.4.1.1 Definição do problema

Antes de iniciar qualquer processo de descoberta, o usuário deve definir exatamente o que ele deseja obter, ou seja os objetivos do processo.

3.4.1.2 Obtenção dos dados

Visa a identificação de informações que possam ser relevantes para o estudo e uma primeira familiarização com seu conteúdo, descrição, qualidade e utilidade. A coleção inicial dos dados procura adquirir a informação com a qual se irá trabalhar, relacionando suas fontes, procedimentos de leitura e os problemas detectados. Nessa tarefa, descreve-se ainda a forma como os dados foram adquiridos, listando seu formato, volume, significado e toda a informação relevante. Durante essa etapa, são realizadas as primeiras descobertas.

3.4.1.3 Seleção dos dados

Todo processo necessita de informações relevantes de entrada. Isto significa que nem todas as informações disponíveis necessitam ser processadas. Deve-se ter o cuidado de selecionar somente aquelas que estejam realmente dentro do contexto dos objetivos. A utilização de informações irrelevantes ou em grande quantidade pode gerar resultados incorretos, além de tornar o processo mais demorado. Consiste em uma série de atividades destinadas a obter o conjunto final de dados, a partir do qual será criado e validado o modelo, (CRISP-DM, 2005).

3.4.1.4 Limpeza dos dados

O objetivo dessa tarefa é remover ruídos e imperfeições. Em textos, seus métodos vão desde a limpeza de caracteres indesejados, correção ortográfica e a normalização do vocabulário (por análise morfológica, *stemming* ou o uso de dicionários).

3.4.1.5 Transformação dos dados

A criação de um conjunto de dados para teste permite construir um mecanismo para comprovar a qualidade e validar os modelos que serão obtidos.

3.4.2 Mineração

São selecionadas e aplicadas técnicas de mineração de dados mais apropriadas, dependendo dos objetivos pretendidos. A mineração representa a fase central do modelo, incluindo escolha, parametrização e execução de técnicas sobre o conjunto de dados visando à criação de um ou vários modelos.

Diferentes técnicas podem ser utilizadas no processo de mineração de dados. Cada uma delas determina uma forma de aplicação e metodologia diferenciada, a utilização ou não de determinada técnica deve ser definida a partir do ambiente trabalhado e do objetivo a ser alcançado. Entre as técnicas aplicadas encontram-se: regressão, associação, *clustering*, classificação, sumarização, detecção de desvios, seqüência e agrupamento por séries temporais.

3.4.3 Pós-processamento

Consiste na revisão dos passos seguidos, verificando se os resultados obtidos vão ao encontro dos objetivos, previamente determinados na definição do problema. De acordo com o resultado alcançados decide-se pela continuidade ou se deverão ser efetuadas correções, voltando às fases anteriores ou ainda, iniciando um novo processo (CRISP-DM, 2005).

4 APLICAÇÃO DO MODELO PROPOSTO PARA TEXT MINING EM SEGURANÇA PÚBLICA

Para o processo de extração de conhecimento dos dados da Secretaria de Segurança Pública e Defesa do Cidadão seguiu-se a metodologia ilustrada na Figura 15 da página 58.

4.1 Pré-processamento

É o primeiro passo do ciclo de mineração de textos, composta de 5 etapas. Durante essa fase realizou-se a implementação do *software* ABC *Clean* v1.0 que lê todos os registros de ocorrência, verificando os erros de ortografia e corrigindo-os automaticamente. A linguagem de programação utilizada para a implementação foi *Delphi* 6.0® da *Borland.*

Como requisitos mínimos dos sistemas é necessário que haja:

a) um ambiente ou sistema operacional gráfico padrão Windows 95;
b) um processador Pentium ou compatível;
c) 32 Mb de memória RAM;
d) 10 Mb de espaço disponível em disco.

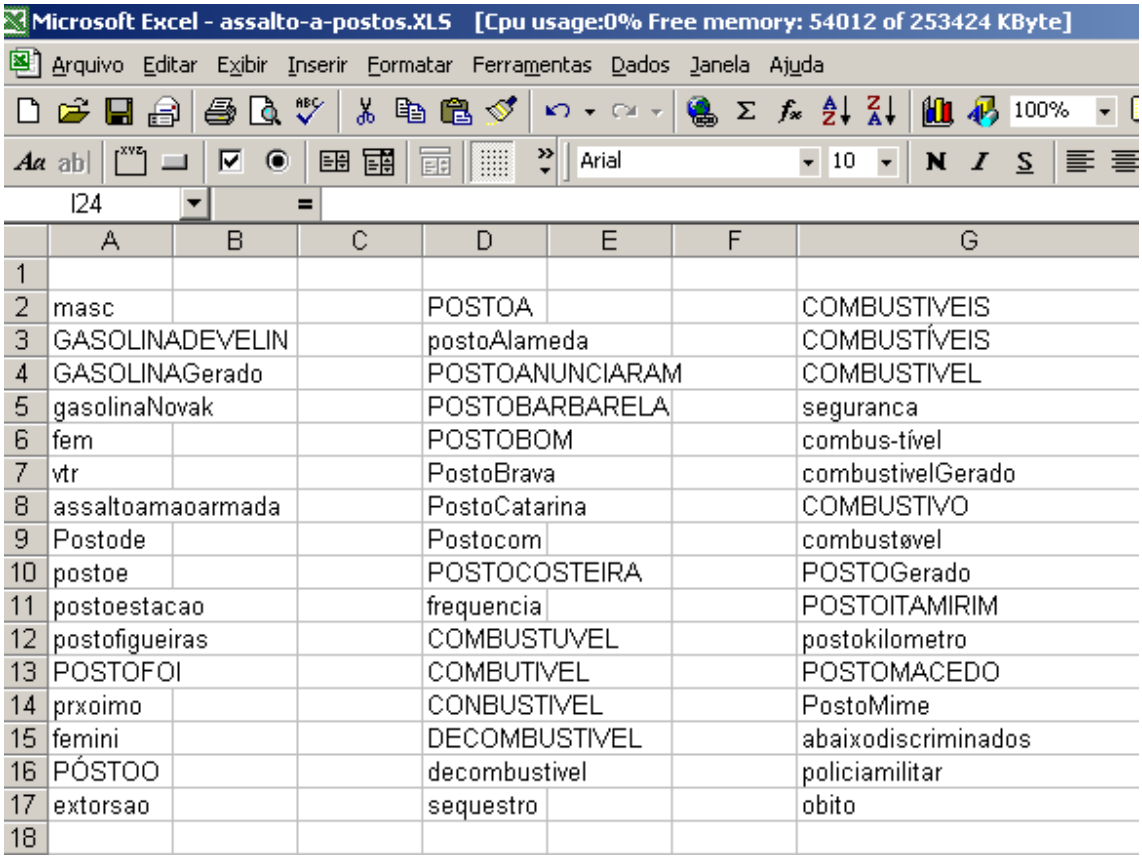
Com o objetivo de aumentar o tempo de resposta do protótipo (levando-se em conta a grande massa de dados a serem processados) recomenda-se 128 Mb de memória RAM.

4.1.1 Problema com os dados

Para uma completa estruturação dos dados da Secretaria de Segurança Pública e Defesa do Cidadão de Santa Catarina, torna-se necessário realizar um tratamento em documentos textuais, em especial, nos registros de ocorrência da polícia civil. Embora parte desses documentos estejam estruturados (como a classificação do tipo de ocorrência, rua, bairro, sexo), a outra parte que contém informações relevantes, não está (descrição dos delitos). Pela ausência dessa estrutura na descrição do texto, esses dados são estudados e trabalhados para a aplicação de *text mining*.

Os dados coletados na Secretaria de Segurança Pública e Defesa do Cidadão de Santa Catarina, por não estarem estruturados apresentaram os seguintes problemas:

a) erros de português (Figura 16), por exemplo: “progeteis, projeties, seguestro, seqestro, tiiros, tirus”;
b) divergência entre natureza de operação e a descrição propriamente dita (Figura 17), por exemplo: o registro de ocorrência 1030271 possui divergência entre a natureza de operação e a descrição, deveria ser D309 - óbito no local e não C903 – comunicação falsa;
c) inadequação dos formatos dos arquivos - os documentos vinham ora em *MsWord*®, ora em *MsExcel*® (Figura 18 e 19);
d) inadequação nos formatos da descrição - as descrições continham caracteres especiais, por exemplo, o *enter* quando utilizado gerava um novo registro de ocorrência, que por sua vez, multiplicava o número de registros (Figura 20).



	A	B	C	D	E	F	G
1							
2	masc			POSTOA			COMBUSTIVEIS
3	GASOLINADEVELIN			postoAlameda			COMBUSTÍVEIS
4	GASOLINAGerado			POSTOANUNCIARAM			COMBUSTIVEL
5	gasolinaNovak			POSTOBARBARELA			seguranca
6	fem			POSTOBOM			combus-tível
7	vtr			PostoBrava			combustivelGerado
8	assaltoamaoarmada			PostoCatarina			COMBUSTIVO
9	Postode			Postocom			combustøvel
10	postoe			POSTOCOSTEIRA			POSTOGerado
11	postoestacao			frequencia			POSTOITAMIRIM
12	postofigueiras			COMBUSTUVEL			postokilometro
13	POSTOFOI			COMBUTIVEL			POSTOMACEDO
14	prxoimo			CONBUSTIVEL			PostoMime
15	femini			DECOMBUSTIVEL			abaixodiscriminados
16	PÓSTOO			decombustivel			policiamilitar
17	extorsao			sequestro			obito
18							

Figura 16 – Erros de português

O registro de ocorrência 1030271 (Figura 17), é um exemplo claro de distorção de informação. A natureza correta desse registro é D309 (óbito no local) e não C903 – Comunicação falsa. Neste caso a digitação incorreta da natureza de operação, gera um registro de ocorrência inconsistente. Ao localizar-se os óbitos, esse registro não será listado e nem contabilizado.

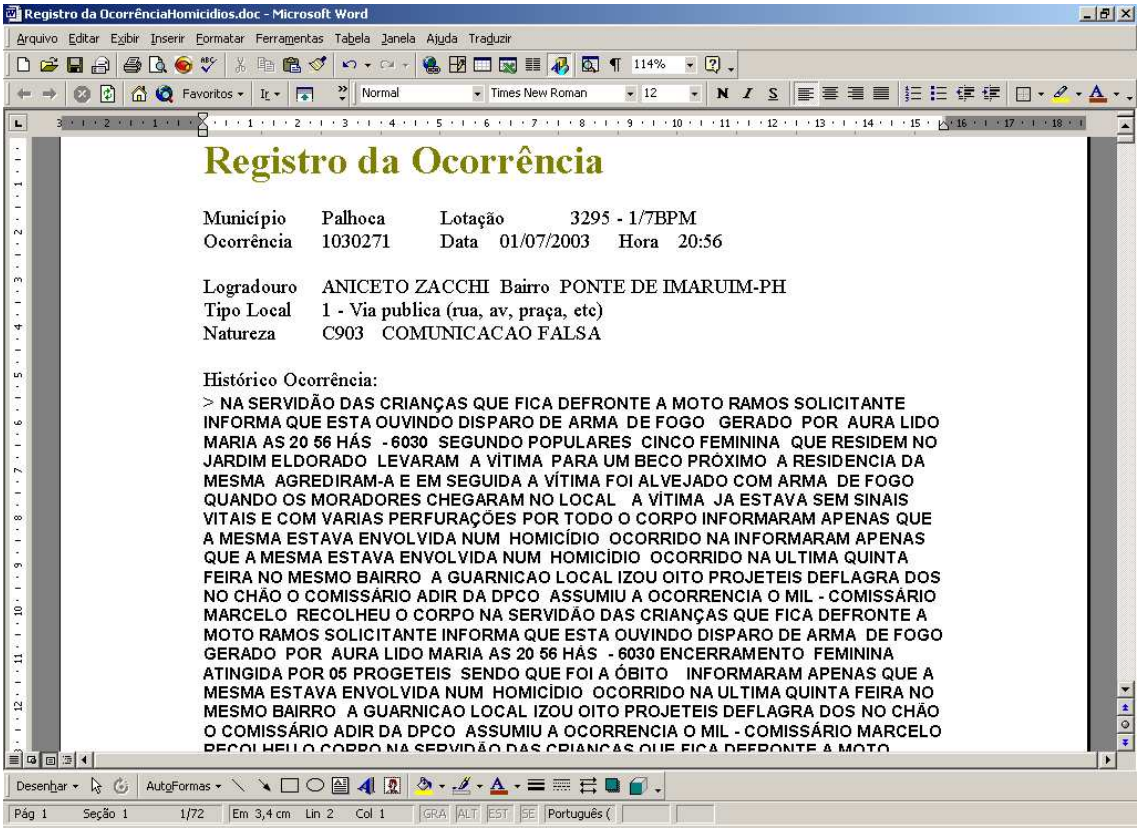


Figura 17 - Divergência entre a natureza de operação e a descrição, deveria ser D309 - óbito no local e não C903 – comunicação falsa.

4.1.2 Obtenção dos dados

Os dados foram obtidos junto a Secretaria de Segurança Pública através da DIRC. Foram cuidadosamente analisados a fim de conhecê-los. Os mesmos foram enviados pela DIRC de várias formas: xls, doc e txt.

4.1.3 Seleção dos dados

O primeiro arquivo (Figura 18) recebido pela DIRC chegou no formato .doc (*MsWord*®) o que dificulta a mineração dos dados. A conversão para *MsExcel*® só foi possível através de uma macro construída no próprio *MsWord*®. Essa macro elimina dados como: município, logradouro, bairro, tipo de local, dados do envolvido tais como idade, sexo, lesão e qualificação, deixando o número da ocorrência, a natureza de

operação e o histórico da ocorrência, que serão utilizados para a mineração. Esses dados que foram eliminados, só foram excluídos desse estudo pois nem todos os arquivos enviados pela DIRC continham essas informações.

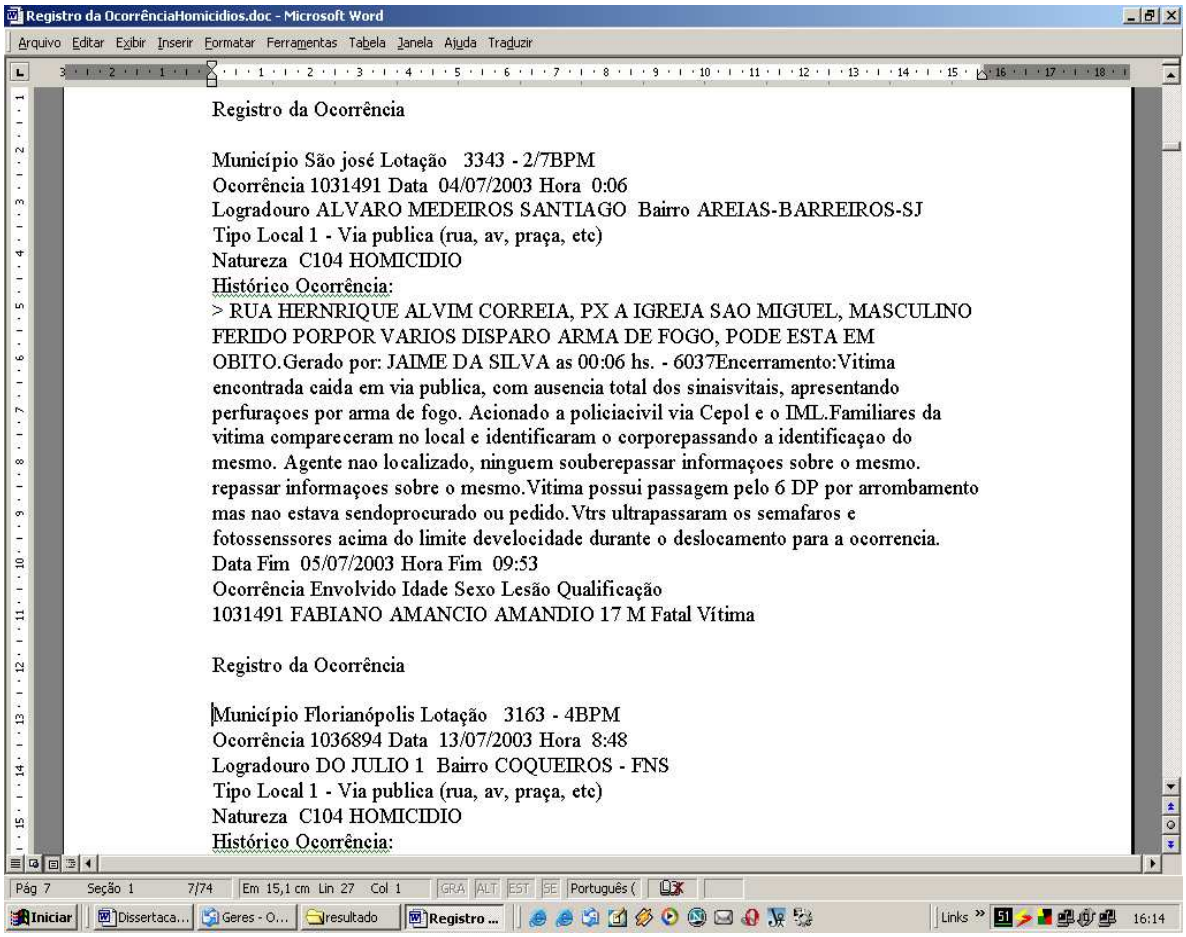
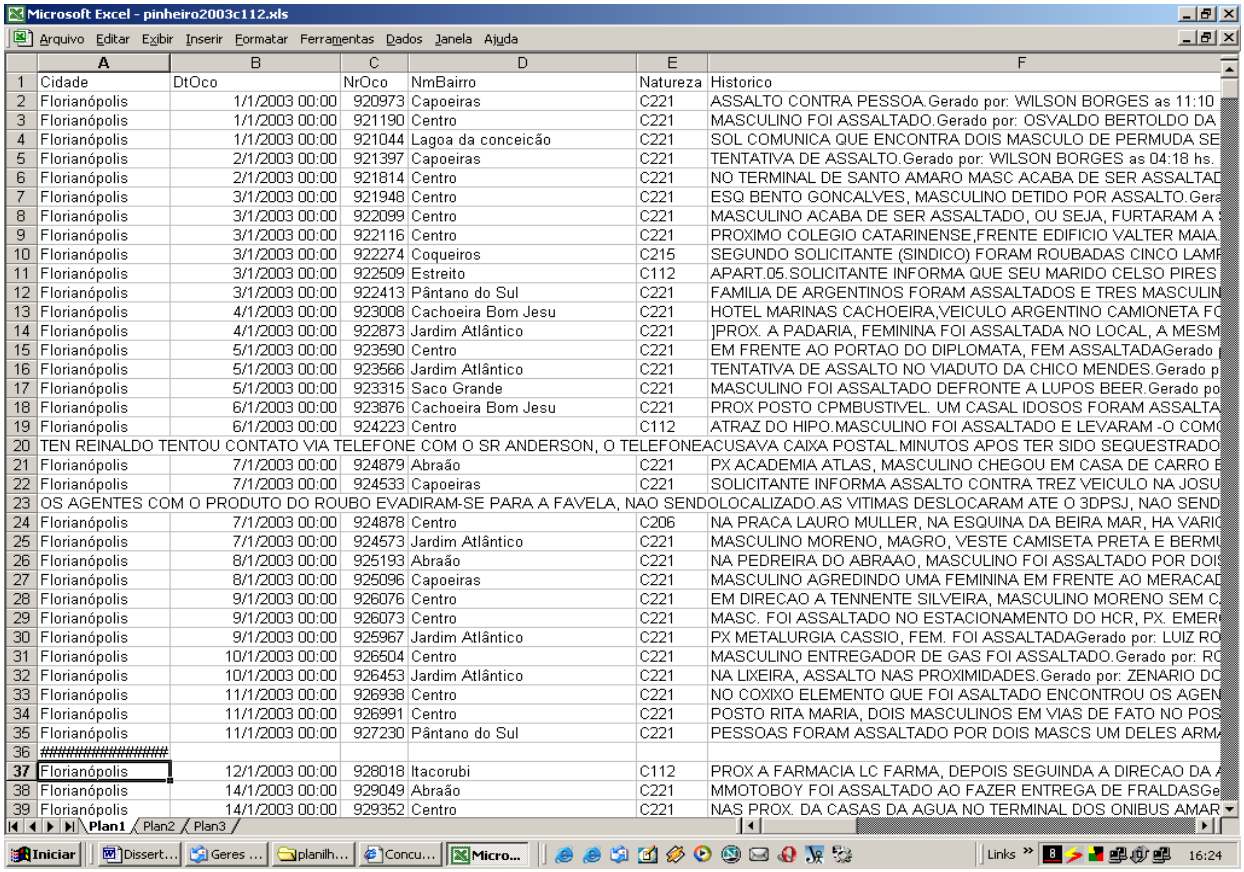


Figura 18 – Registros de homicídios no formato .doc

O arquivo representado na Figura 19, foi o segundo arquivo enviado pela DIRC. Ele apresentava um problema grave, alguns históricos vinham com campos estourados, como pode ser visto na linha 36.



Microsoft Excel - pinheiro2003c112.xls

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

	A	B	C	D	E	F
1	Cidade	DtOco	NrOco	NmBairro	Natureza	Historico
2	Florianópolis	1/1/2003 00:00	920973	Capoeiras	C221	ASSALTO CONTRA PESSOA.Gerado por: WILSON BORGES as 11:10
3	Florianópolis	1/1/2003 00:00	921190	Centro	C221	MASCULINO FOI ASSALTADO.Gerado por: OSVALDO BERTOLDO DA
4	Florianópolis	1/1/2003 00:00	921044	Lagoa da conceicão	C221	SOL COMUNICA QUE ENCONTRA DOIS MASCULO DE PERMUDA SE
5	Florianópolis	2/1/2003 00:00	921397	Capoeiras	C221	TENTATIVA DE ASSALTO.Gerado por: WILSON BORGES as 04:18 hs.
6	Florianópolis	2/1/2003 00:00	921814	Centro	C221	NO TERMINAL DE SANTO AMARO MASC ACABA DE SER ASSALTAD
7	Florianópolis	3/1/2003 00:00	921948	Centro	C221	ESQ BENTO GONCALVES, MASCULINO DETIDO POR ASSALTO.Gera
8	Florianópolis	3/1/2003 00:00	922099	Centro	C221	MASCULINO ACABA DE SER ASSALTADO, OU SEJA, FURTARAM A
9	Florianópolis	3/1/2003 00:00	922116	Centro	C221	PROXIMO COLEGIO CATARINENSE,FRENTE EDIFICIO VALTER MAIA.
10	Florianópolis	3/1/2003 00:00	922274	Coqueiros	C215	SEGUNDO SOLICITANTE (SINDICO) FORAM ROUBADAS CINCO LAMP
11	Florianópolis	3/1/2003 00:00	922509	Estreito	C112	APART.05.SOLICITANTE INFORMA QUE SEU MARIDO CELSO PIRES
12	Florianópolis	3/1/2003 00:00	922413	Pântano do Sul	C221	FAMILIA DE ARGENTINOS FORAM ASSALTADOS E TRES MASCULIN
13	Florianópolis	4/1/2003 00:00	923008	Cachoeira Bom Jesu	C221	HOTEL MARINAS CACHOEIRA,VEICULO ARGENTINO CAMIONETA FO
14	Florianópolis	4/1/2003 00:00	922873	Jardim Atlântico	C221	]PROX. A PADARIA, FEMININA FOI ASSALTADA NO LOCAL, A MESM
15	Florianópolis	5/1/2003 00:00	923590	Centro	C221	EM FRENTE AO PORTAO DO DIPLOMATA, FEM ASSALTADAGerado
16	Florianópolis	5/1/2003 00:00	923566	Jardim Atlântico	C221	TENTATIVA DE ASSALTO NO VIADUTO DA CHICO MENDES.Gerado p
17	Florianópolis	5/1/2003 00:00	923315	Saco Grande	C221	MASCULINO FOI ASSALTADO DEFRONTE A LUPOS BEER.Gerado po
18	Florianópolis	6/1/2003 00:00	923876	Cachoeira Bom Jesu	C221	PROX POSTO CPMBUSTIVEL. UM CASAL IDOSOS FORAM ASSALTA
19	Florianópolis	6/1/2003 00:00	924223	Centro	C112	ATRAZ DO HIPO.MASCULINO FOI ASSALTADO E LEVARAM -O COM
20	TEN REINALDO TENTOU CONTATO VIA TELEFONE COM O SR ANDERSON, O TELEFONEACUSAVA CAIXA POSTAL MINUTOS APOS TER SIDO SEQUESTRADO					
21	Florianópolis	7/1/2003 00:00	924879	Abraão	C221	PX ACADEMIA ATLAS, MASCULINO CHEGOU EM CASA DE CARRO E
22	Florianópolis	7/1/2003 00:00	924533	Capoeiras	C221	SOLICITANTE INFORMA ASSALTO CONTRA TREZ VEICULO NA JOSU
23	OS AGENTES COM O PRODUTO DO ROUBO EVADIRAM-SE PARA A FAVELA, NAO SENDOLOCALIZADO.AS VITIMAS DESLOCARAM ATE O 3DPSJ, NAO SEND					
24	Florianópolis	7/1/2003 00:00	924878	Centro	C206	NA PRACA LAURO MULLER, NA ESQUINA DA BEIRA MAR, HA VARIC
25	Florianópolis	7/1/2003 00:00	924573	Jardim Atlântico	C221	MASCULINO MORENO, MAGRO, VESTE CAMISETA PRETA E BERMU
26	Florianópolis	8/1/2003 00:00	925193	Abraão	C221	NA PEDREIRA DO ABRAAO, MASCULINO FOI ASSALTADO POR DOIS
27	Florianópolis	8/1/2003 00:00	925096	Capoeiras	C221	MASCULINO AGREDINDO UMA FEMININA EM FRENTE AO MERACAD
28	Florianópolis	9/1/2003 00:00	926076	Centro	C221	EM DIRECAO A TENNENTE SILVEIRA, MASCULINO MORENO SEM C
29	Florianópolis	9/1/2003 00:00	926073	Centro	C221	MASC. FOI ASSALTADO NO ESTACIONAMENTO DO HCR, PX. EMER
30	Florianópolis	9/1/2003 00:00	925967	Jardim Atlântico	C221	PX METALURGIA CASSIO, FEM. FOI ASSALTADAGerado por: LUIZ RO
31	Florianópolis	10/1/2003 00:00	926504	Centro	C221	MASCULINO ENTREGADOR DE GAS FOI ASSALTADO.Gerado por: RO
32	Florianópolis	10/1/2003 00:00	926453	Jardim Atlântico	C221	NA LIXEIRA, ASSALTO NAS PROXIMIDADES.Gerado por: ZENARIO DO
33	Florianópolis	11/1/2003 00:00	926938	Centro	C221	NO COXIXO ELEMENTO QUE FOI ASALTADO ENCONTROU OS AGEN
34	Florianópolis	11/1/2003 00:00	926991	Centro	C221	POSTO RITA MARIA, DOIS MASCULINOS EM VIAS DE FATO NO POS
35	Florianópolis	11/1/2003 00:00	927230	Pântano do Sul	C221	PESSOAS FORAM ASSALTADO POR DOIS MASCS UM DELES ARMA
36	################					
37	Florianópolis	12/1/2003 00:00	928018	Itacorubi	C112	PROX A FARMACIA LC FARMA, DEPOIS SEGUINDA A DIRECAO DA
38	Florianópolis	14/1/2003 00:00	929049	Abraão	C221	MMOTOBOY FOI ASSALTADO AO FAZER ENTREGA DE FRALDASGe
39	Florianópolis	14/1/2003 00:00	929352	Centro	C221	NAS PROX. DA CASAS DA AGUA NO TERMINAL DOS ONIBUS AMAR

Plan1 Plan2 Plan3

Iniciar | Dissert... | Geres ... | planilh... | Concu... | Micro... | Links » | 16:24

Figura 19 – Formato inadequado arquivo .xls, com campos estourados

Os caracteres especiais, tais como TAB e ENTER geram registros novos ao serem lidos pelo *software* ABC *Mining* de carga de dados. Os arquivos enviados pela DIRC vinham com esses caracteres inseridos dentro do histórico do registro de ocorrência. Na Figura 20, pode-se ver claramente esse caracter □.

```
Florianópolis     2003-01-01 00:00:00     920973      Capoeiras
      C221  ASSALTO CONTRA PESSOA.Gerado por: WILSON BORGES as
11:10 hs. - 6026Encerramento: VITIMA QUANDO CAMINHAVA NA VILA
APARECIDA, AVISTOU AGENTEQUE ROUBOU SUA BICICLETA.NESTE INSTANTE,
A VTR 2684 PASSAVA PELO LOCAL EFOI ACIONADA PARA DETER O AGENTE.NO
MOMENTO O AGENTE NAO ESTAVA COM A BICICLETA, POREM, LEVOU A
GUARNIÇAO EAVITIMA NO .LOCAL ONDE HAVIA NEGOCIADO A MESMA.A
BICICLETA FOI RECUPERADA NO BECO DO HELIO NA FAVELA CHICO
MENDES.DADOS DA BICICLETA: CALOI DE 21 MARCHA. AGENTE QUE FURTOU,
AGENTE QUE
DADOS DA BICICLETA: CALOI DE 21 MARCHA. AGENTE QUE FURTOU, AGENTE
QUECOMPROU O FURTO E VITIMA, FORAM ENCAMINHADOS PARA A 9ª
DELEGACIA DEPOLICIA.
```

Florianópolis 2003-01-01 00:00:00 920973 Capoeiras C221 ASSALTO CONTRA PESSOA.Gerado por: WILSON BORGES as 11:10 hs. - 6026Encerramento: VITIMA QUANDO CAMINHAVA NA VILA APARECIDA, AVISTOU AGENTEQUE ROUBOU SUA BICICLETA.NESTE INSTANTE, A VTR 2684 PASSAVA PELO LOCAL EFOI ACIONADA PARA DETER O AGENTE.NO MOMENTO O AGENTE NAO ESTAVA COM A BICICLETA, POREM, LEVOU A GUARNIÇAO EAVITIMA NO .LOCAL ONDE HAVIA NEGOCIADO A MESMA.A BICICLETA FOI RECUPERADA NO BECO DO HELIO NA FAVELA CHICO MENDES.DADOS DA BICICLETA: CALOI DE 21 MARCHA. AGENTE QUE FURTOU, AGENTE QUE
DADOS DA BICICLETA: CALOI DE 21 MARCHA. AGENTE QUE FURTOU, AGENTE QUECOMPROU O FURTO E VITIMA, FORAM ENCAMINHADOS PARA A 9ª DELEGACIA DEPOLICIA.

Figura 20 – Exemplo de ocorrência que continha caracteres especiais

Quanto à inadequação dos formatos elaborou-se um formato padrão. Para esse estudo, criou-se um arquivo no formato .txt, através dos seguintes passos:

a) o arquivo do *MsWord*® (Registro da OcorrênciaHomicidios.doc), que continha a natureza de operação C104[7], foi convertido para .xls (Excel);

b) o arquivo do *MsExcel*® (pinheiro2003c112.xls), que continha as naturezas C112, C207, C221 e C215[8] permaneceu no formato que estava;

c) os dois arquivos, que agora estão em .xls foram agrupados em um outro arquivo chamado ocorrencias.xls onde todas as ocorrências formaram uma única base de dados;

d) esse arquivo foi importado no *MsAccess*®, e exportado para o formato .txt delimitado por *pipe* "|", conforme demonstra a Figura 21.

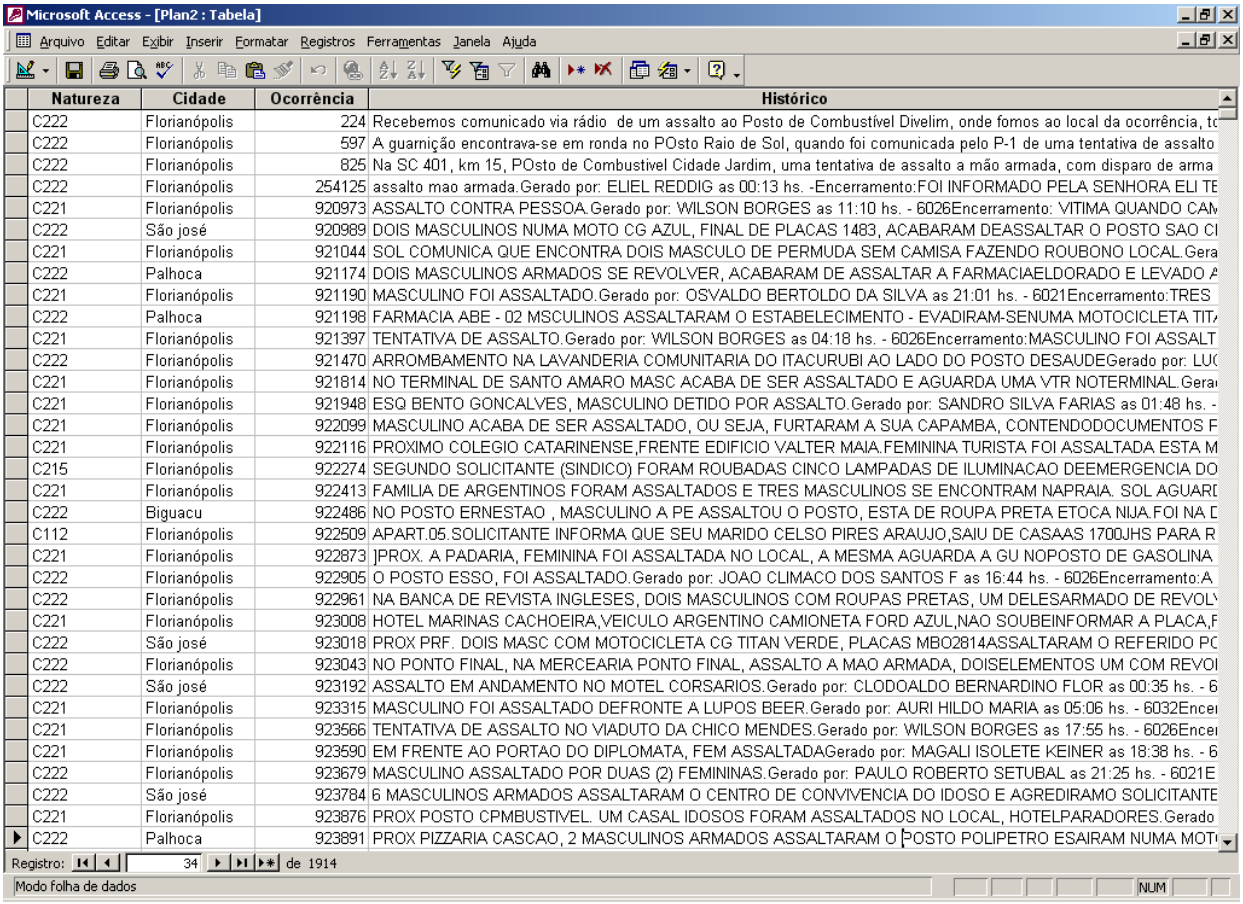


Figura 21 - Dados importados para o *MsAccess*®

[7] Registros policias do tipo Homicídio.

[8] Para saber qual o delito cometido para essa ocorrência consulte o anexo A.

Dessa forma o *software* ABC *Clean* desenvolvido pôde importá-los para limpeza, conforme demonstra a Figura 22:

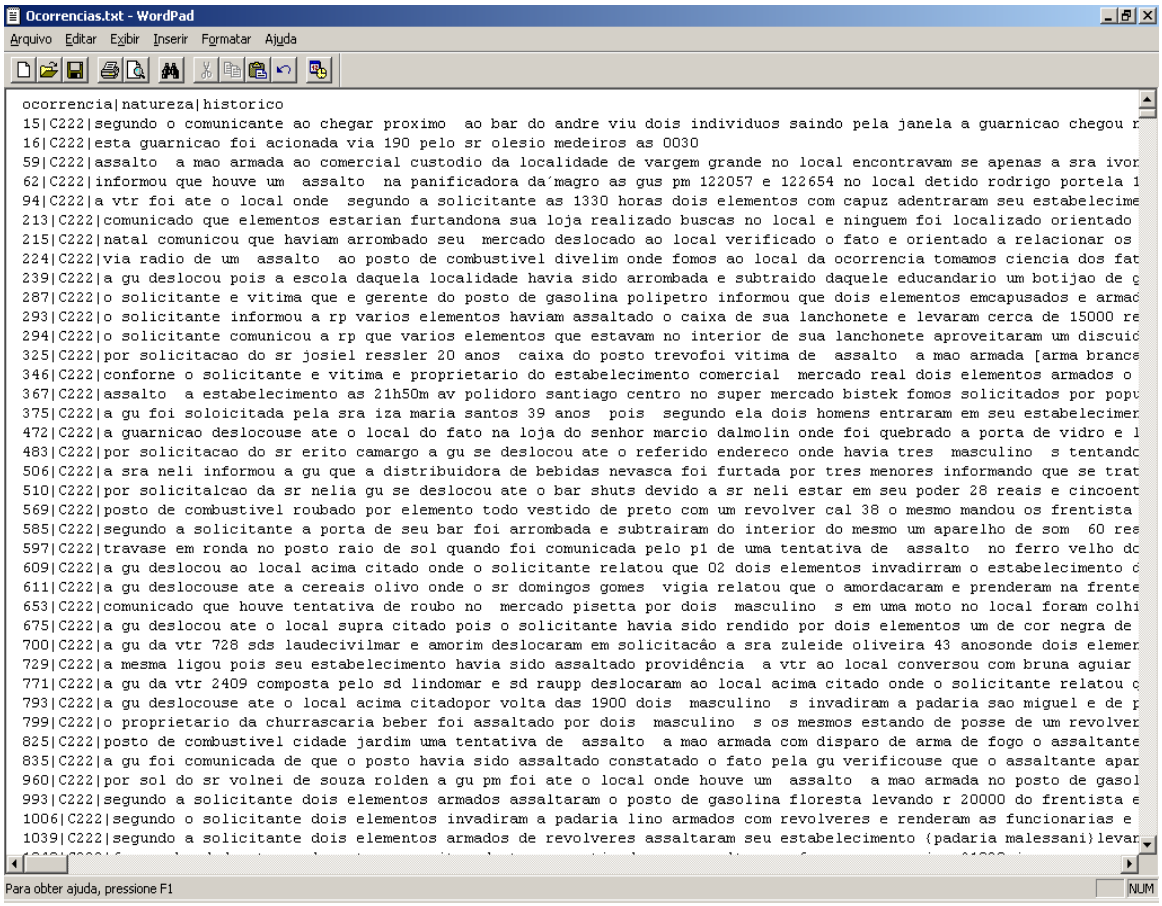


Figura 22 - Dados prontos para utilização do ABC *Clean*

Recomenda-se que os dados brutos enviados (para o arquivo a ser minerado) sofram um processo de limpeza dos caracteres especiais, excluindo-os da descrição e que os próximos arquivos sejam exportados do banco de dados da Polícia Militar no formato .txt delimitado por | (*pipe*).

Desses registros, 121 ocorrências policiais estão classificadas corretamente e serviram como base de treinamento para a construção da árvore de decisão. No anexo C, estão listados esses ROs.

4.1.4 Limpeza dos dados

Os erros de português devem ser corrigidos para não distorcer a freqüência das palavras. Para isso criou-se a ferramenta ABC *Clean*. O ABC *Clean* faz a normalização do vocabulário, com a ajuda do dicionário *Aspell*. A ferramenta foi construída na língua portuguesa “do Brasil”, por ser inédita neste idioma.

4.1.4.1 ABC *Clean*

O elevado número de erros de ortografia encontrados na base, fez com que houvesse a necessidade de implementação desse *software*, já que não existia nenhum que fizesse a limpeza para a língua portuguesa falada no Brasil. O diagrama de atividades na Figura 23 ajuda o entendimento do funcionamento do *software* de limpeza da base ABC *Clean*. Para o desenvolvimento do diagrama de atividades utilizou-se a ferramenta *Jude*® da *Eiwa System Management Inc*.

Para usar o ABC Clean, selecione os arquivos que deverão estar no sub-diretório IN. Se este sub-diretório não existir no mesmo diretório em que se encontra o ABC *Clean*, o mesmo será criado automaticamente. Note que é possível remover alguns caracteres especiais desses arquivos como o CR (*Carriage Return* ou *Enter*) e o LF (*Line Feed*). A Figura 24, demonstra a tela principal do *software* ABC *Clean*.

Todos os arquivos que estiverem no sub-diretório IN aparecem na relação de arquivos para limpeza. Após ter marcado os arquivos para serem corrigidos, o próximo passo consiste na verificação da ortografia, conforme a Figura 25. Nesta etapa o *software* lê cada arquivo texto e vai montando uma lista com todas as palavras erradas.

Informa ao ABC *Clean* que ao encontrar palavras (erradas) que pertençam ao mesmo grupo, substituir aquelas que tiverem peso menor por aquela que tiver o maior peso (Figura 26). Tanto o grupo quanto o peso são informados pelo usuário. A identificação de semelhanças é opcional, o usuário pode ir direto a etapa 4 se assim desejar (Figura 27). No exemplo da Figura 26, para determinar-se que as palavras prx, prox, próx e próxim pertencem ao mesmo grupo, determinou-se que o grupo era o de número 70 (poderia ser qualquer número) e que a palavra com maior peso seria próxim, que recebeu o valor 100, identificando que a palavra com maior peso era essa, para esse

grupo. Toda vez que for encontrado prx, prox, próx e próxim o *software* ABC *Clean* aprendeu que elas devem ser substituídas pela palavra próxim e essa por sua vez por próximo.

Substitui-se as palavras erradas pelas corretas contidas no dicionário *Aspell* (veja exemplo na Figura 28). Caso o *Aspell* não tenha conseguido achar uma palavra equivalente no dicionário, o usuário poderá cadastrar a palavra não encontrada na opção Palavra Nova, no menu de opções da etapa 4.

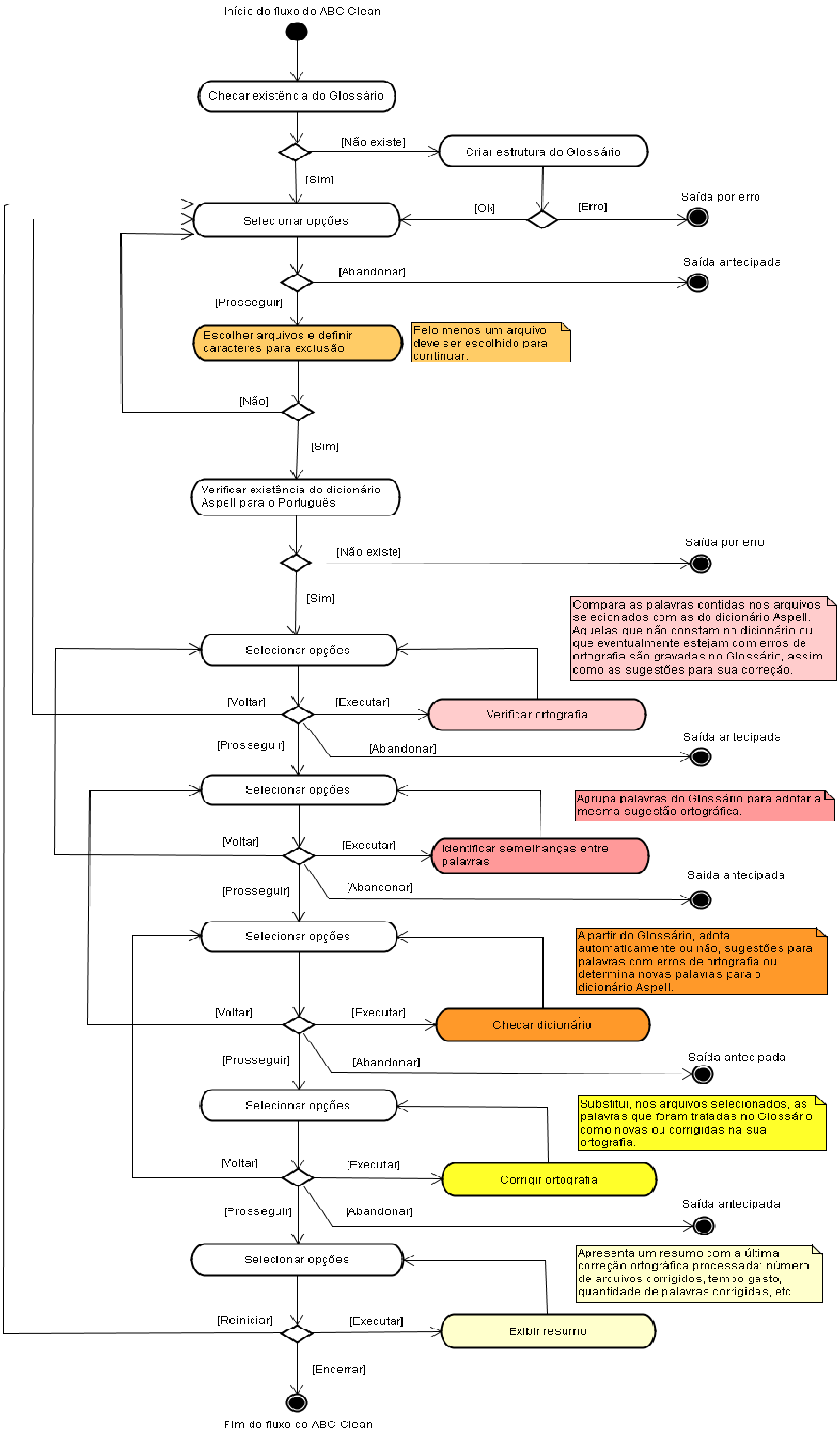


Figura 23 - Diagrama de atividades para a construção do *software* ABC *Clean*

Figura 24 - Interface do *software ABC Clean*, construído para a limpeza da base

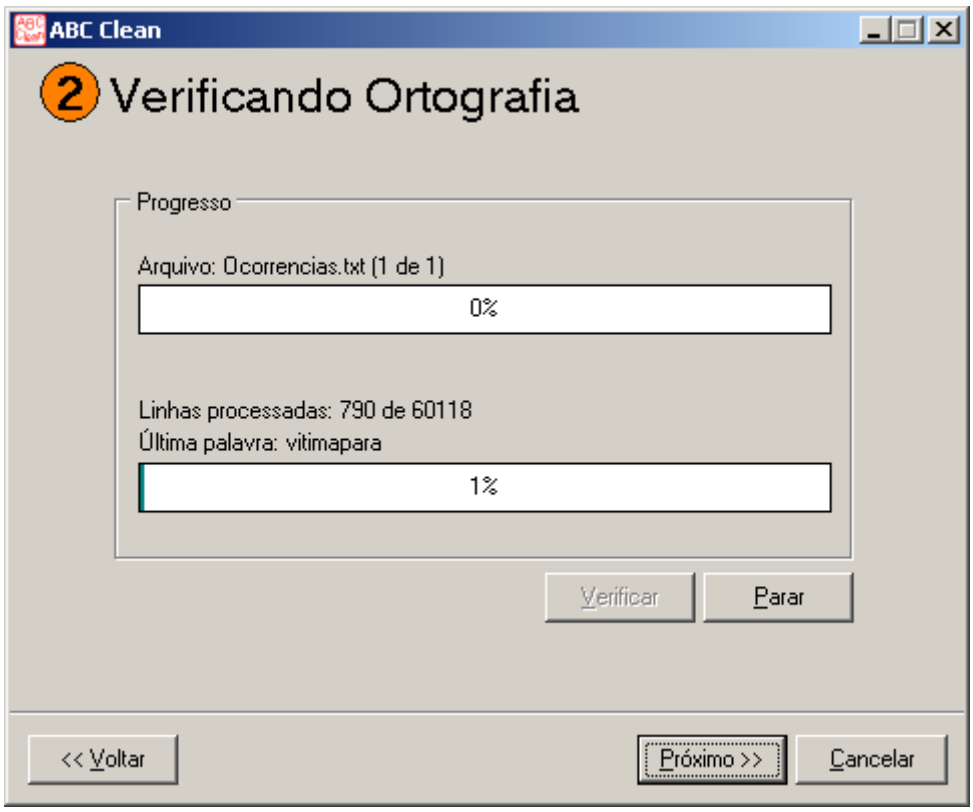


Figura 25 - Montando lista com todas as palavras erradas.

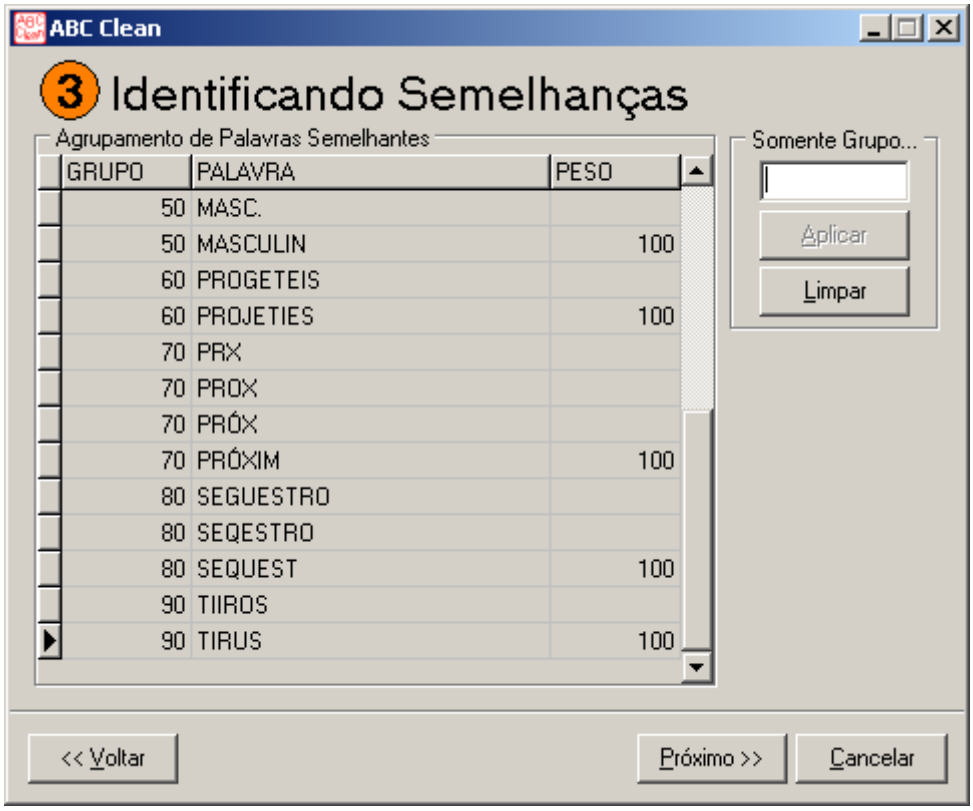


Figura 26 – Identificação de semelhanças

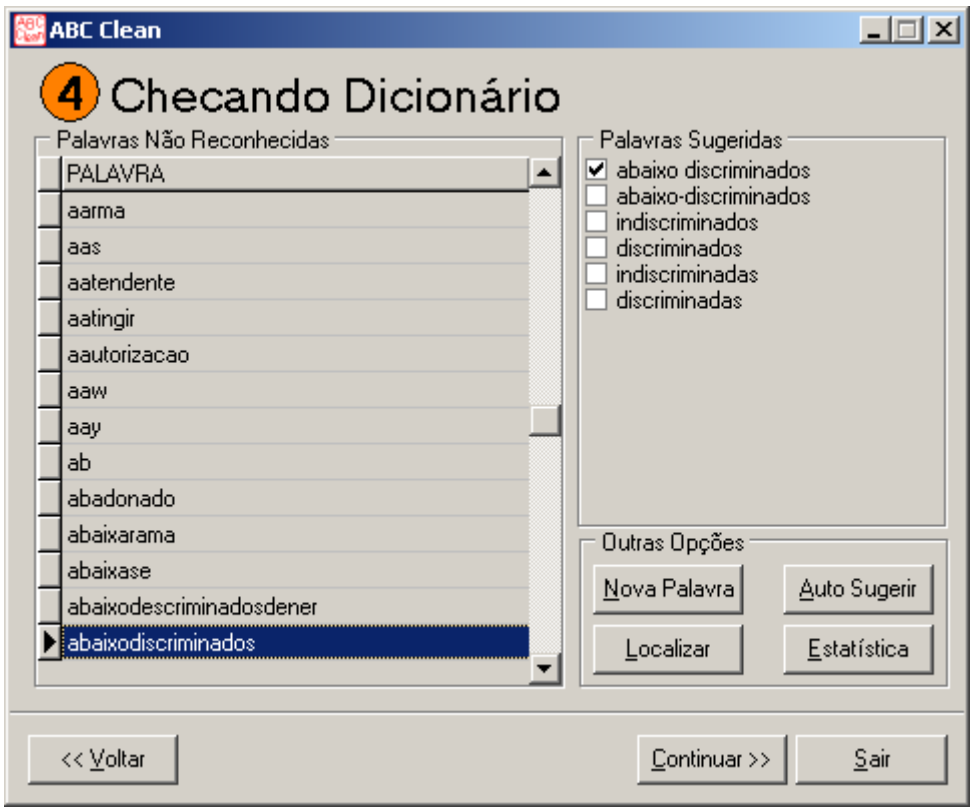


Figura 27 – Checagem do dicionário *Aspell*

Figura 28 – Resultados gerados pelo *ABC Clean*

Na etapa 5, é realizada a correção ortográfica e geração dos arquivos de textos limpos. Esses arquivos são gerados no subdiretório OUT, Figura 29.

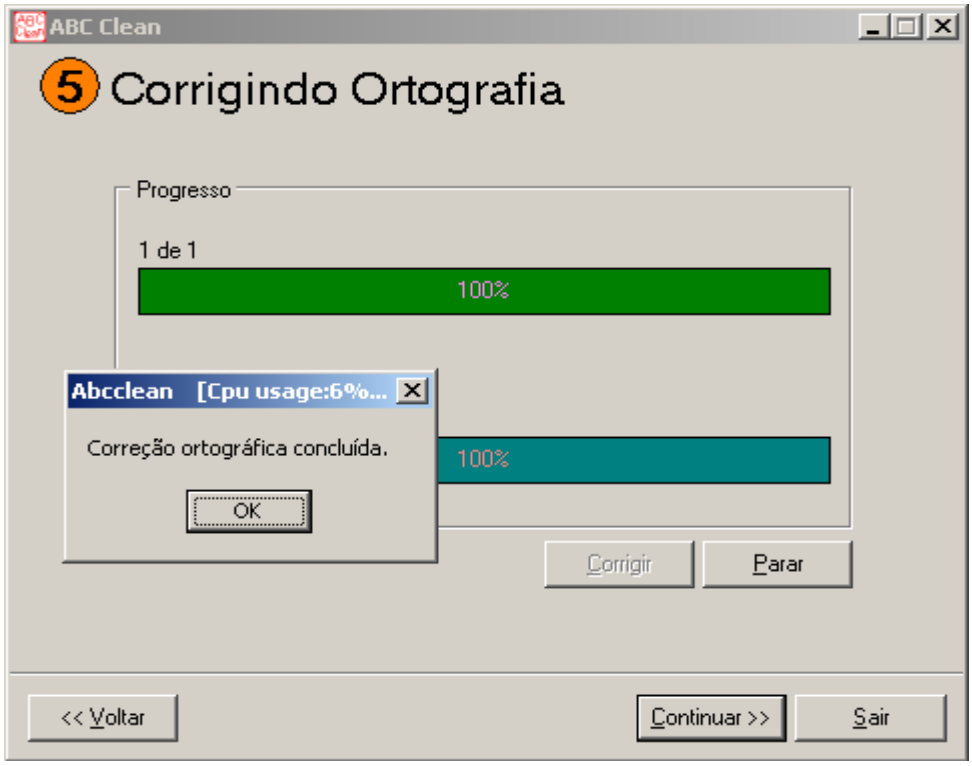


Figura 29 – Correção da ortografia e geração de arquivos limpos

Na etapa 6 é exibido um resumo do que foi feito (Figura 30).

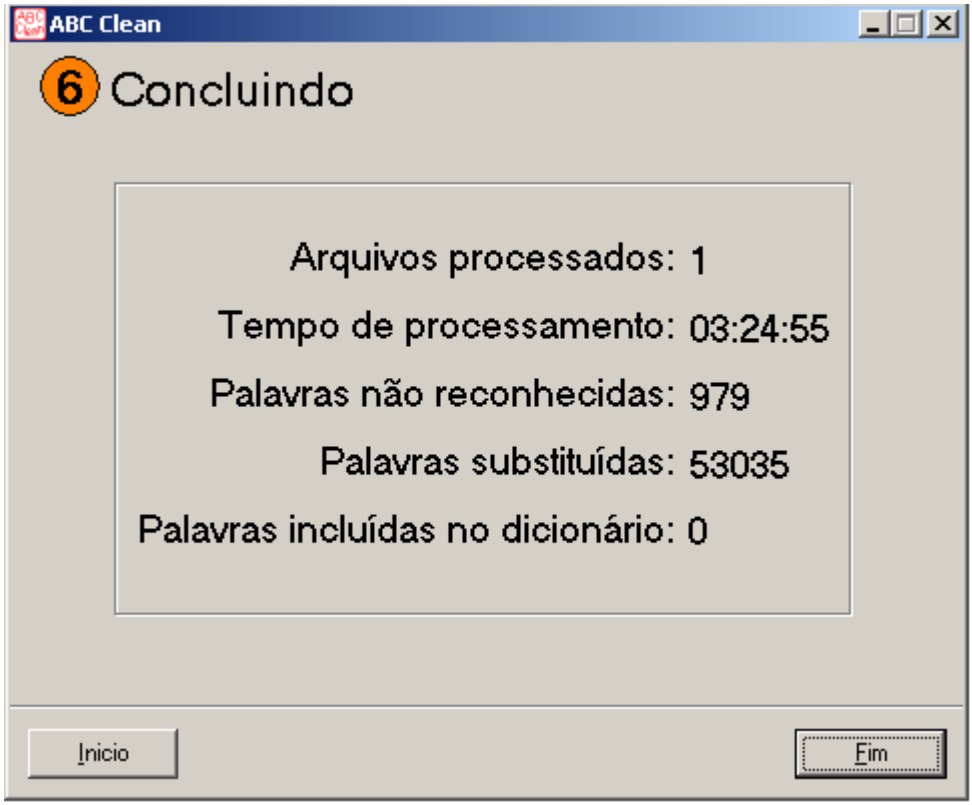


Figura 30 - Resumo do que foi feito

Foram efetuadas várias correções pelo ABC *Clean* nos RO enviados pela DIRC. Por exemplo: “informa**co**es” por “informa**ção**es”, “abaixodiscriminados” por “abaixo discriminados”, “irregularidade constatad**o”** por “irregularidade constatad**a”**.

A parte de limpeza dos dados foi uma tarefa árdua, pois não existem na literatura trabalhos semelhantes.

4.2 Mineração

Caracteriza-se pela transformação dos dados tratados em conhecimento. Para isso emprega tecnologia conhecida como *text mining*, que tem a finalidade de realizar a exploração e análise de dados por meio automático ou semi-automático, em busca de relacionamentos entre dados, padrões, regras que caracterizam tendências. Nesta etapa

aplicam-se os algoritmos de mineração. Nessa dissertação aplicou-se o algoritmo ID3 com o auxílio do *Weka*® para a geração das regras de classificação e do *software* ABC *Mining* para reclassificação dos ROs, geração do arquivo para o *Weka*® minerar, e demais opções descritas abaixo.

4.2.1 ABC *Mining*

Depois de corrigido os dados e, portanto, com uma base limpa, passou-se à fase de mineração de dados com o uso do *software* ABC *Mining* v1.0, Figura 31. A interface da ferramenta é constituída de 8 opções:

a) carga de dados;
b) *stopwords;*
c) categorias
d) *keywords;*
e) exportação para o *Weka®;*
f) classificação dos registros;
g) pesquisar;
h) sair.

O ABC *Mining* v1.0 lê todos os ROs e gera arquivos no formato texto para auxiliar o *Weka*® na geração da árvore de decisão, permite o cadastramento e visualização das *keywords*, *stopwords* e categorias. A opção de pesquisar também é uma ferramenta importante para localizar rapidamente os registros, já que permite localizar uma palavra, frase ou *stopword*, *keyword* pela sua freqüência.

Após o *Weka*® gerar as regras, o ABC *Mining* reclassifica os registros de ocorrência, apontando sua nova natureza de operação. Alguns registros não serão reclassificados, por não ter sido utilizada uma base com todos os tipos de natureza de operação.

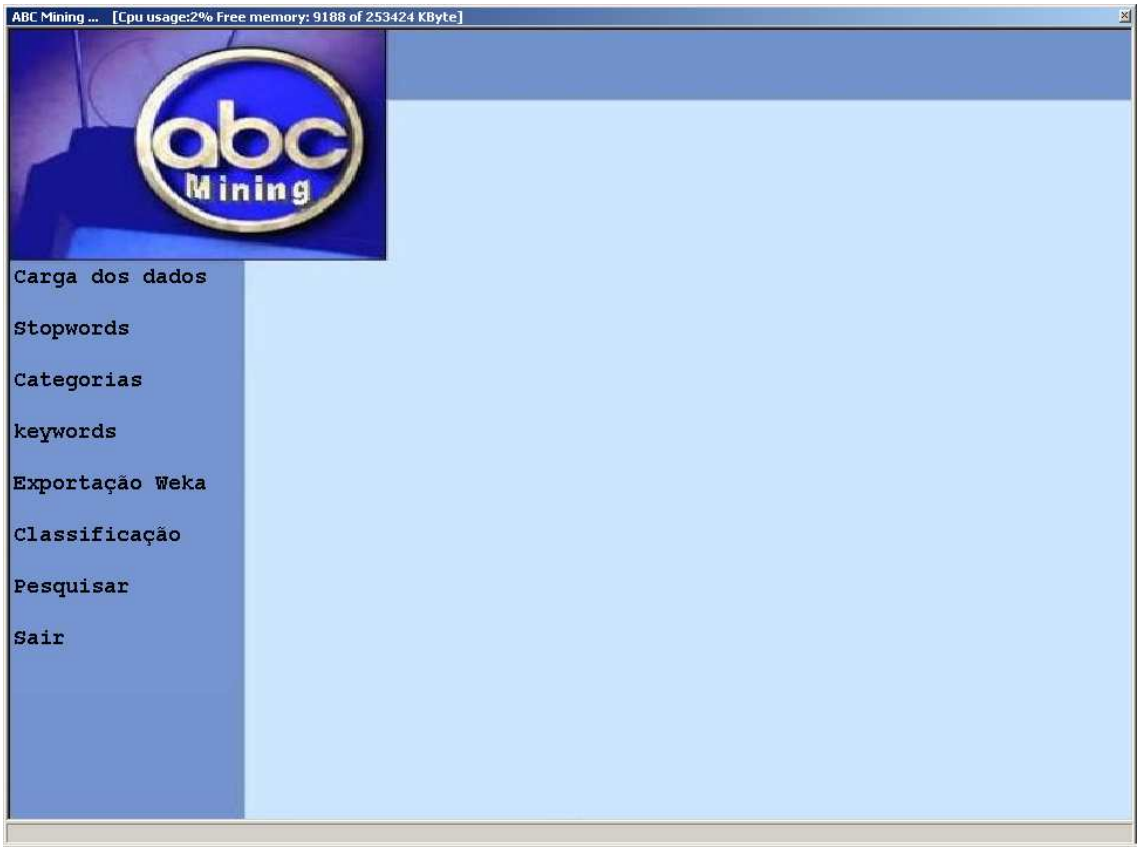


Figura 31 – Interface do *software* de mineração ABC *Mining*

4.2.1.1 Carga de dados

O primeiro passo para utilização do sistema consiste na carga dos dados, Figura 32. Todos os registros de ocorrência são carregados para a memória do computador. Nesse momento são gerados os vetores de palavras, Figura 33. Cada palavra é agrupada pela sua ortografia, por isso é necessário que as palavras estejam escritas corretamente.

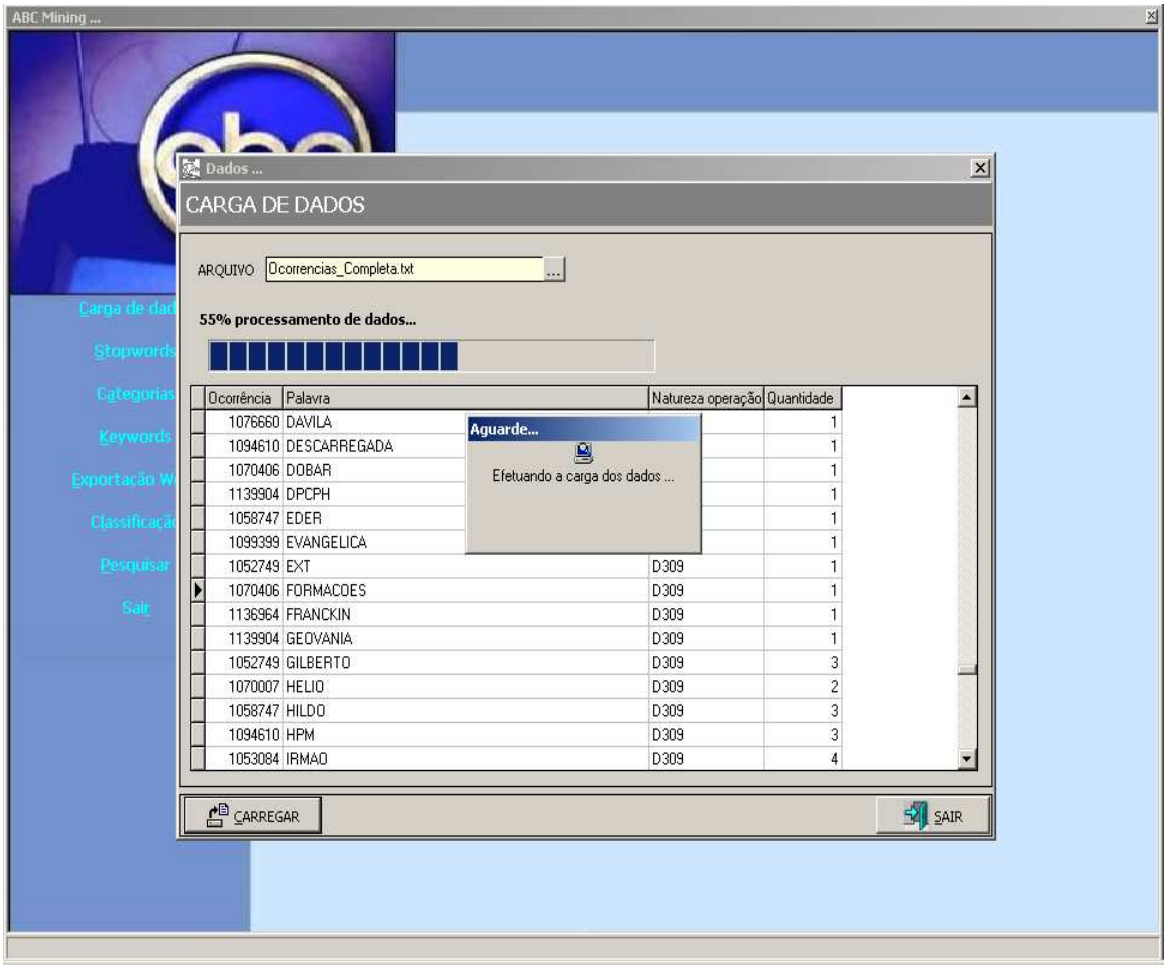


Figura 32 – Carga de dados

Na carga dos dados, verificou-se alguns resultados, conforme a Tabela 5:

Tabela 5 - Totais obtidos na carga de dados

Palavras	Total
Lidas	23.996.148
Lidas não repetidas	17354
Stopwords	23.950.637
Stopwords não repetidas	16775
Utilizadas na mineração	45511
Utilizadas na mineração não repetidas	579

Para carregar os ROs na memória o ABC *Mining* levou 0:30s.

Palavra	Quantidade
ASSALTO	3385
ARMA	3141
REVOLVER	1661
MOTO	1296
DINHEIRO	1275
COR	1159
AGENTE	1156
RONDAS	1030
ESTABELECIMENTO	929
POSTO	918
MERCADO	783
CAIXA	774
CAPACETE	759
CALCA	746
ROUBAR	741
EVADIRAM	652
MORENO	651
ORIENTADO	648
CAMISA	639
JAQUETA	627
PLACA	615
BUSCAS	602
LEVARAM	553
CONTATO	543
EFETUARAM	474
TRAJANDO	463
LOJA	448
FOGO	406
DELEGACIA	398

Figura 33 – Vetor com as 29 palavras mais freqüentes desconsiderando-se as *stopwords*

4.2.1.2 Stopwords

A lista de s*topwords* possui 16.775 únicas, totalizando 23.950.637 *stopwords* encontradas. O algoritmo varre todo o texto à procura das palavras contidas nesta lista. Quando localizada, estas palavras são removidas para que não participem da análise. A Figura 34 demonstra algumas delas, as palavras que estão à esquerda são as *stopwords* e as palavras à direita são as palavras-chave, com sua respectiva quantidade e freqüência relativa.

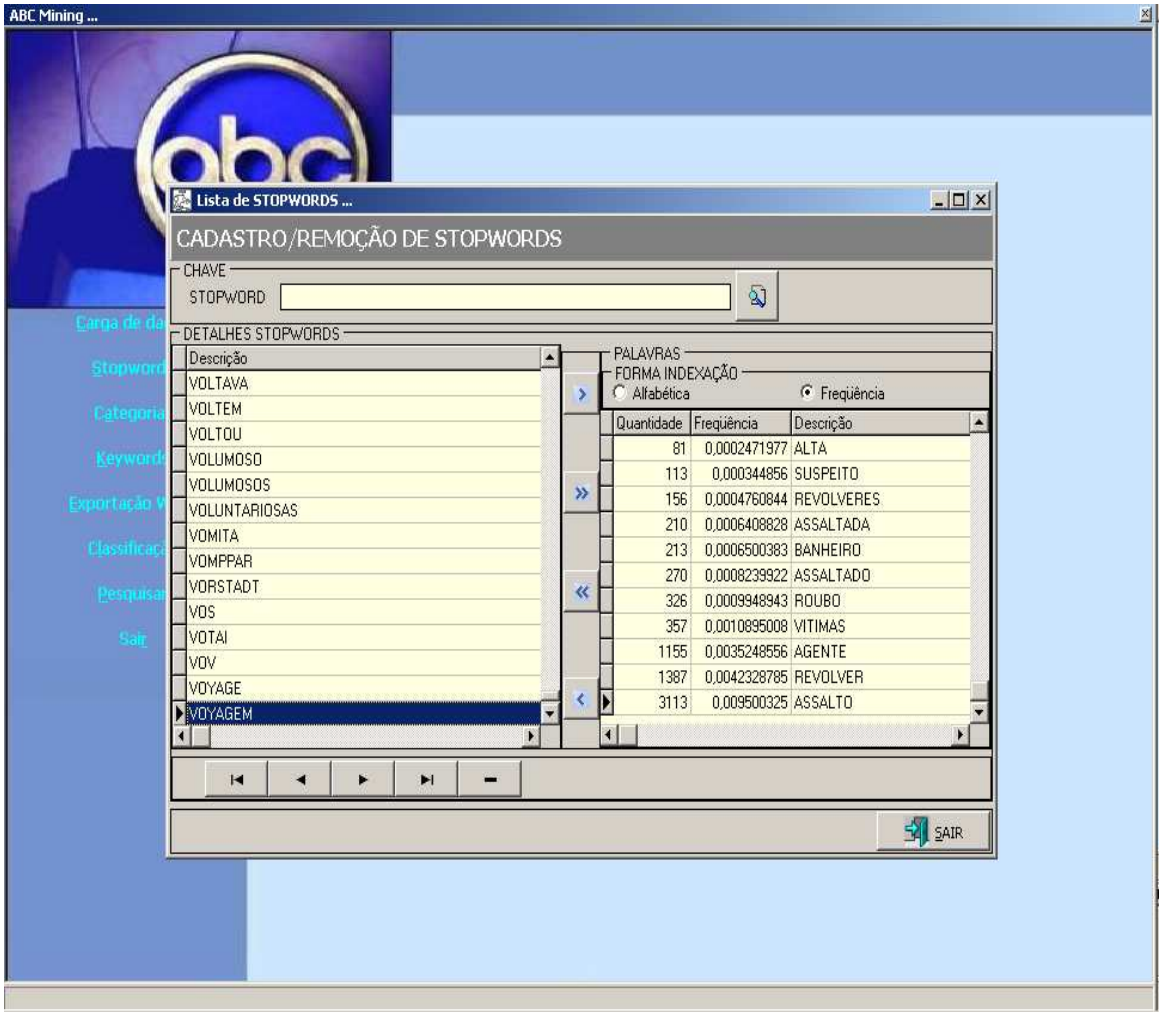


Figura 34 – Lista de *stopwords*

4.2.1.3 Categorias

As categorias (naturezas de operações), devem ser cadastradas, possibilitando a análise e geração dos arquivos para que o *software Weka*® gere a árvore de decisão. Quando informada uma categoria ao sistema, este verifica a existência ou não. Caso não exista, ele permite o cadastramento. A Figura 35 demonstra todas as categorias enviadas pela DIRC.

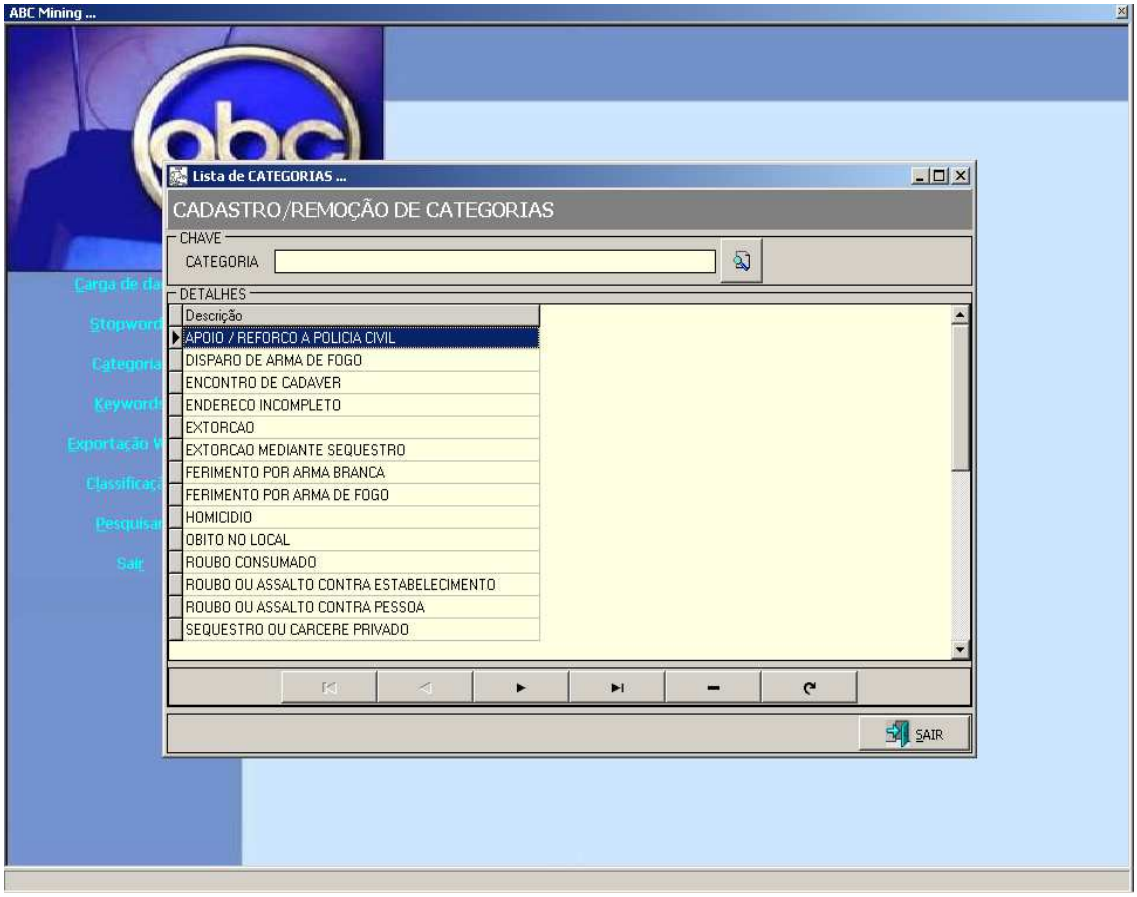


Figura 35 - Definição das categorias

4.2.1.4 Keywords

Para encontrar as palavras-chave para cada natureza de operação, utilizou-se a lista de palavras lidas na carga de dados menos as *stopwords*. Na carga de dados, também foi carregado para a memória, o *corpus* de referência, que é composto por vários documentos utilizados como fonte para esta pesquisa. A diferença entre o *corpus* de referência e o *corpus* de estudo, permite encontrar as palavras mais freqüêntes. Porém, como se utilizam várias naturezas de operação foi necessário encontrar as palavras mais freqüentes para cada categoria. Desta forma, atribui-se uma ou várias *keywords* para cada categoria (Figura 36).

Figura 36 – *Keywords* encontradas pelo sistema para cada natureza de operação

Estando-se em constante contato com a base de dados, sendo minerando ou mesmo lendo os ROs, verificou-se que as palavras-chaves encontradas para cada natureza de operação estudada representam as palavras que sempre estão incluídas nos textos da descrição do RO. A partir dessas palavras pode-se gerar o arquivo para que o *software Weka*® gere a árvore de decisão.

4.2.1.5 Exportação para o *Weka*®

Após definir quais são as palavras-chave (*Keywords*) para cada natureza de operação, é necessário descobrir quais palavras aparecem ou não em cada registro de ocorrência. Então, foi gerado um arquivo no formato ARFF para que o *Weka*® possa importá-lo e minerá-lo. O arquivo é composto das seguintes partes:

a) @ relation ocorrencia: define o nome do arquivo como sendo ocorrencia;
b) @ attribute: o código da natureza de operação e cada palavra-chave são definidos como atributo. Cada atributo pode assumir os valores 0 (não) ou 1 (sim);
c) @ data: indica o início dos dados dos registros de ocorrência;

d) para cada registro de ocorrência é gerado uma linha de dados. Nela é indicado se o atributo existe ou não (0 ou 1).

A Figura 37 exibe o arquivo gerado.

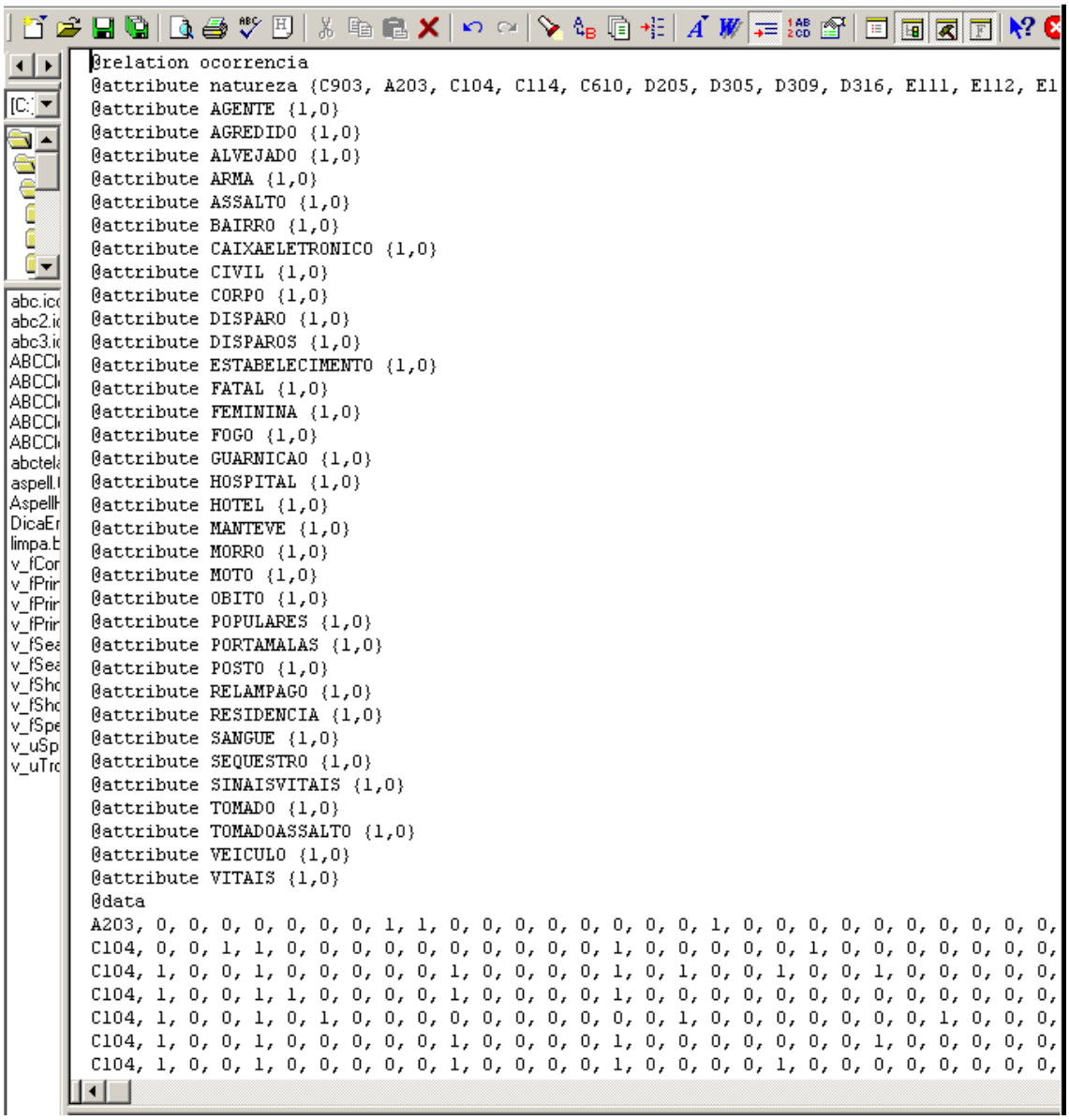


Figura 37 – Arquivo gerado para a mineração com o *Weka*®

4.2.2 Regras geradas pelo *Weka*®

Partindo-se do arquivo ocorrencia.arff gerado pelo ABC *Mining* pode-se gerar a árvore de decisão com o auxílio do *software Weka*®. Usou-se para isso o algoritmo ID3 criado por Quilan. Dentre os resultados obtidos pode-se citar:

a) foram analisadas 34 variáveis (*Keywords*);
b) o tempo gasto para gerar o modelo foi de 1.06 segundos;
c) foram geradas 217 regras.

Da árvore de decisão gerada pelo *Weka*® pode-se visualizar um trecho na Figura 38. Para ver toda a árvore de decisão gerada consulte o apêndice B. Foram utilizados para gerar o arquivo ocorrencia.arff os 121 ROs enviados pela DIRC, onde todos os ROs estavam classificados corretamente.

```
VITAIS = 1
| CORPO = 1
| | SANGUE = 1
| | | BAIRRO = 1: D205
| | | BAIRRO = 0: D305
| | SANGUE = 0
| | | CIVIL = 1
| | | | AGENTE = 1
| | | | | DISPARO = 1: C104
| | | | | DISPARO = 0: E112
| | | | AGENTE = 0: C114
| | | CIVIL = 0
| | | | ALVEJADO = 1
| | | | | DISPARO = 1: D309
| | | | | DISPARO = 0: C104
| | | | ALVEJADO = 0: C104
```

Figura 38 - Parte da árvore de decisão gerada pelo *Weka*®

Da árvore gerada, tem-se também outros dados tais como: sumário, detalhes por classe (taxa de erro, precisão, etc) e matriz de referência cruzada.

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2538	94,5608 %
Incorrectly Classified Instance	146	5,4396 %
Kappa statistic	0.3652	
Total Number of Instances	2684	

Segundo a Tabela 2 da página 46 para a medida *Kappa* = 0,3652 determina que a concordância é média/baixa.

No caso da natureza de operação C222, pode-se observar na próxima página, que a taxa de acerto é de 98,2%, sendo que uma grande parte dos registros nessa natureza estavam cadastrados corretamente. A quantidade de exemplos positivos que foram corretamente classificados (*precision*) é alta 0,975, o *recall* = 0,982, a medida F = 0,979 (*F-measure*).

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0	0	0	0	0	A203
0.208	0.006	0.25	0.208	0.227	C104
0	0.002	0	0	0	C112
0.2	0.003	0.125	0.2	0.154	C114
0	0	0	0	0	C207
0	0.001	0	0	0	C215
0.111	0.011	0.125	0.111	0.118	C221
0.982	0.504	0.975	0.982	0.979	C222
0	0.001	0	0	0	C610
0	0	0	0	0	C903
0	0.001	0	0	0	D205
0.5	0.001	0.25	0.5	0.333	D305
0.619	0.002	0.684	0.619	0.65	D309
0	0	0	0	0	D316
0	0	0	0	0	E111
0.3	0.002	0.375	0.3	0.333	E112
0	0.001	0	0	0	E123

=== Confusion Matrix ===

	A203	C104	C112	C114	C207	C215	C221	C222	C610	C903	D205	D305	D309	D316	E111	E112	E123	Total
A203	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1[9]
C104	1	5	0	0	0	0	2	6	1	1	0	0	4	1	1	2	0	24
C112	0	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	9
C114	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	5
C207	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	4
C215	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2
C221	0	0	0	0	1	0	4	31	0	0	0	0	0	0	0	0	0	36

[9] Total de ROs por natureza de operação

C222	0	4	5	5	0	2	24	2511	0	0	1	1	0	0	0	3	1	2557
C610	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2
C903	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
D205	0	1	0	0	0	0	0	2	0	0	0	1	0	0	0	0	0	4
D305	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
D309	0	5	0	0	0	0	0	2	1	0	0	0	13	0	0	0	0	21
D316	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	3
E111	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2
E112	0	2	0	1	0	0	0	4	0	0	0	0	0	0	0	3	0	10
E123	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1

Dos 2557 registros com natureza de operação C222, 98,2% estavam com natureza de operação corretas equivalendo 2511 ROs, os 46 restantes, segundo a matriz de referência cruzada, conforme a Tabela 6, estão assim distribuídos:

Tabela 6 - Distribuição dos 46 ROs a serem reclassificados

Natureza de Operação	RO a serem reclassificados
C104	4
C112	5
C114	5
C215	2
C221	24
D205	1
D305	1
E112	3
E123	1

Os valores *precision, recall* e *f-measure* encontrados pelo *Weka*® foram calculados baseando-se na matriz de referência cruzada da página anterior:

precision = 2511 / (2511 + 66) = 0,975

recall = 2511 / (2511 + 46) = 0,982

f-measure = $\frac{2 * 0,975 * 0,982}{0,975 + 0,982} = \frac{1,9149}{1,957} = 0,979$

Na Figura 39, pode-se visualizar um dos gráficos gerados pelo *Weka*®. No eixo X estão representadas as naturezas de operações, no eixo Y tem-se arma (1) e sem arma (0). Em cinza é assalto e em branco é não assalto. Pode-se perceber com o auxílio desse gráfico que as naturezas C221, C222 e C215, são roubos ou assaltos (muito azul) com arma (1).

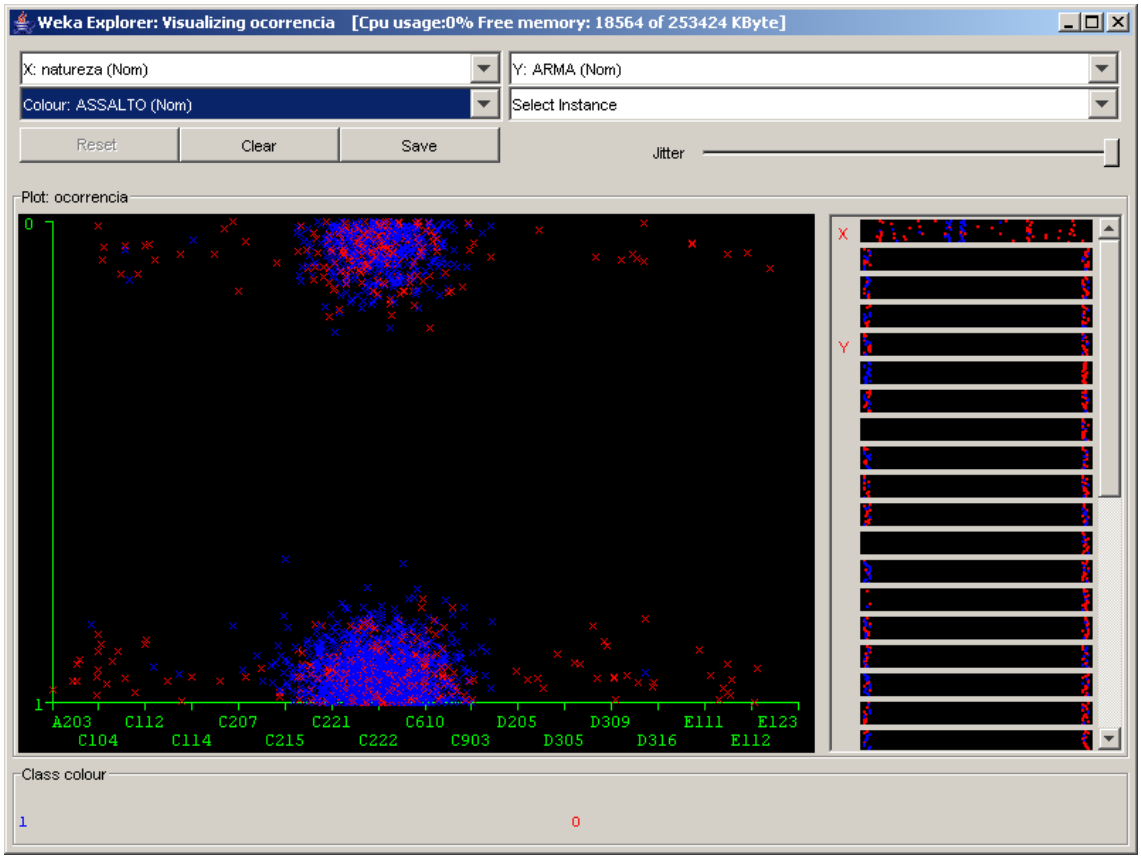


Figura 39 – Gráfico gerado pelo *Weka*® com as palavras ASSALTO e ARMA

Devido a dificuldade de interpretação da árvore gerada pelo *Weka*®, foi necessário a implementação de um *software*, o *ABC Transform*, que interpreta as regras geradas e transforma em um código fonte para ser usado com a linguagem de programação Pascal ou *Delphi*®. Esse *software* pode interpretar qualquer árvore gerada pelo *Weka*® e transformar em regras do tipo *IF THEN (*Figura 40*).*

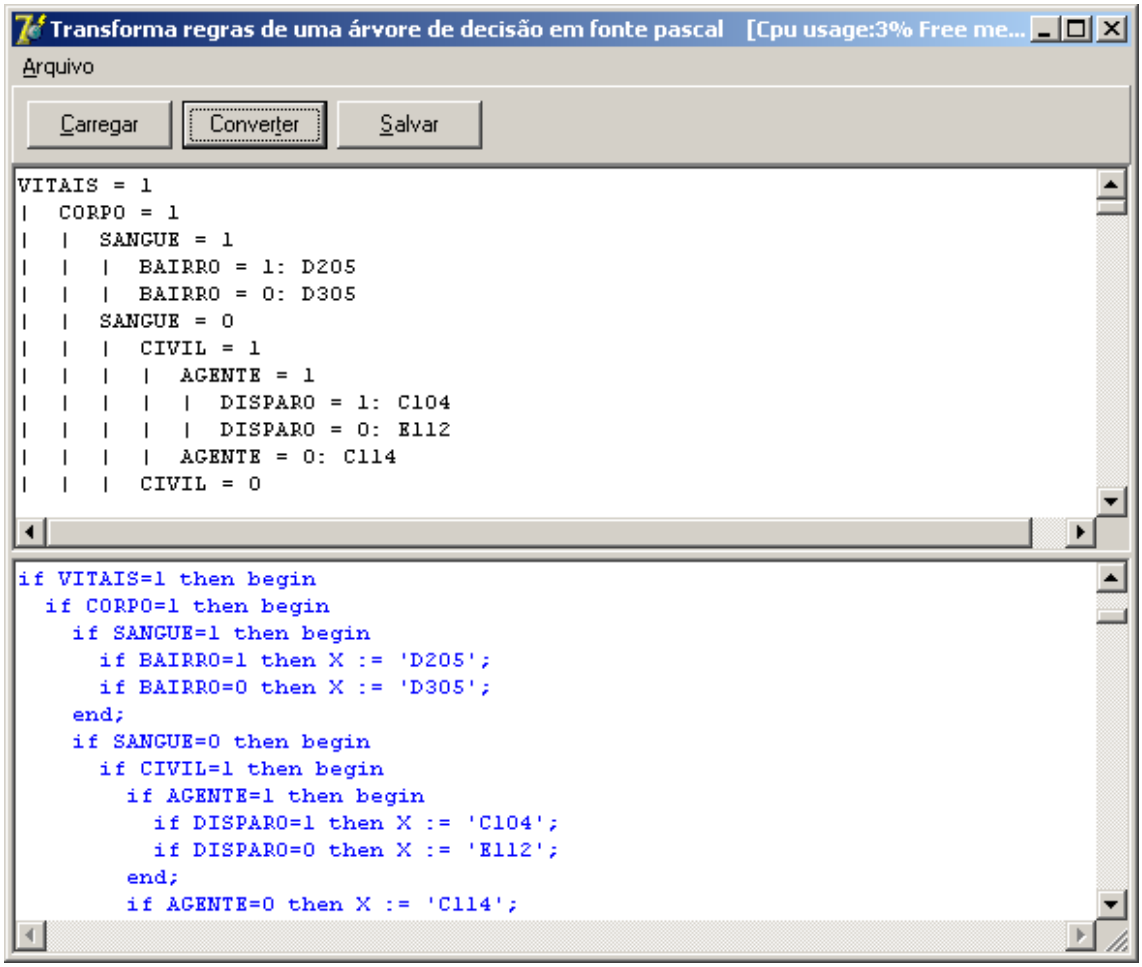


Figura 40 – Tela principal do *ABC Transform*

4.3 Pós-processamento

Nesta etapa os padrões obtidos, ou seja, as regras, são interpretados e avaliados, consistindo na consolidação do conhecimento obtido, possibilitando a associação do mesmo a um sistema facilitando sua utilização em trabalhos futuros.

A avaliação dos dados envolve a comparação entre os resultados previstos e os resultados encontrados sendo possível determinar se a informação resultante pode ser considerada válida e útil.

4.3.1 Reclassificação dos registros de ocorrência

Após as regras terem sido geradas pelo *Weka*®, a partir dos 121 ROs com a natureza de operação correta enviados pela DIRC, aplicou-se essas regras sobre os registros de ocorrência a fim de reclassificá-los. Nesta fase, os registros de ocorrência que podem estar com a natureza de operação incorreta são identificados. Ficando a critério de um especialista em Segurança Pública a avaliação da mudança em definitivo. Com a geração das regras de decisão, tanto os registros existentes podem ser reclassificados como os novos podem ser classificados no momento da sua inclusão na base de dados da Polícia Militar. A Figura 41, ilustra o registro de ocorrência 1030271, que foi citado como exemplo na Figura 17 da página 64, depois de minerado, foi reclassificado para a natureza de operação correta D309 – óbito no local.

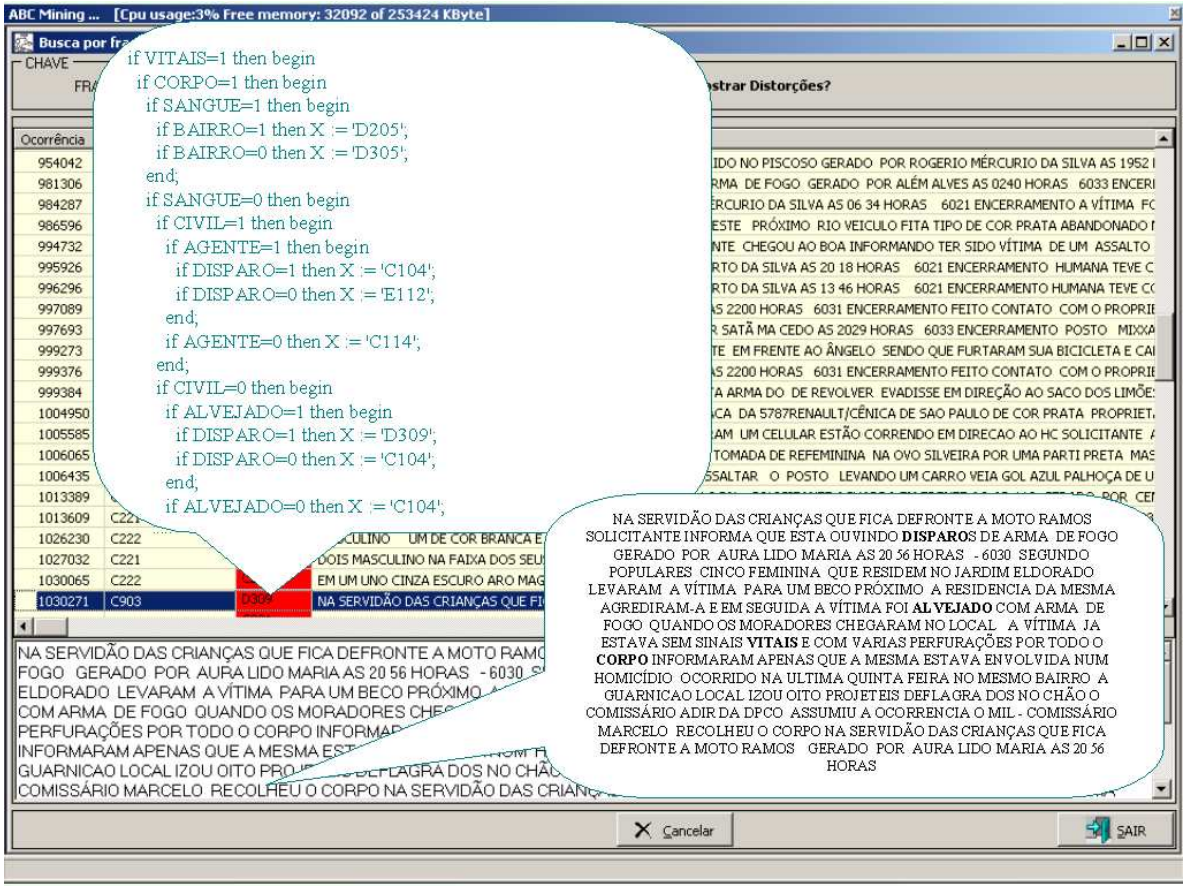


Figura 41 – Registros de ocorrência 1030271 com distorção entre natureza de operação cadastrada e minerada.

Dos 2684 registros de ocorrência minerados, 146 registros encontravam-se com natureza de operação incorretas, totalizando 5,4396% dos registros. O erro de classificação mostrou-se baixo, porém deve-se levar em consideração os seguintes pontos:

a) tem-se uma base de dados bem pequena em relação ao volume dos dados armazenados pela Polícia Militar em seu banco de dados, já que a DIRC atende 4000 chamados por dia;
b) esses ROs foram previamente classificados pelo pessoal da DIRC e esperava-se que estivessem corretos.

Para efeito de teste, substituiu-se, no arquivo ocorrencias-completa.txt, com todos os 2684 registros de ocorrência, a natureza de operação C222 – roubo ou assalto a estabelecimento por C100 (poderia ser qualquer valor), e o modelo reclassificou-os como sendo C222 (Figura 42), validando assim a sua eficiência.

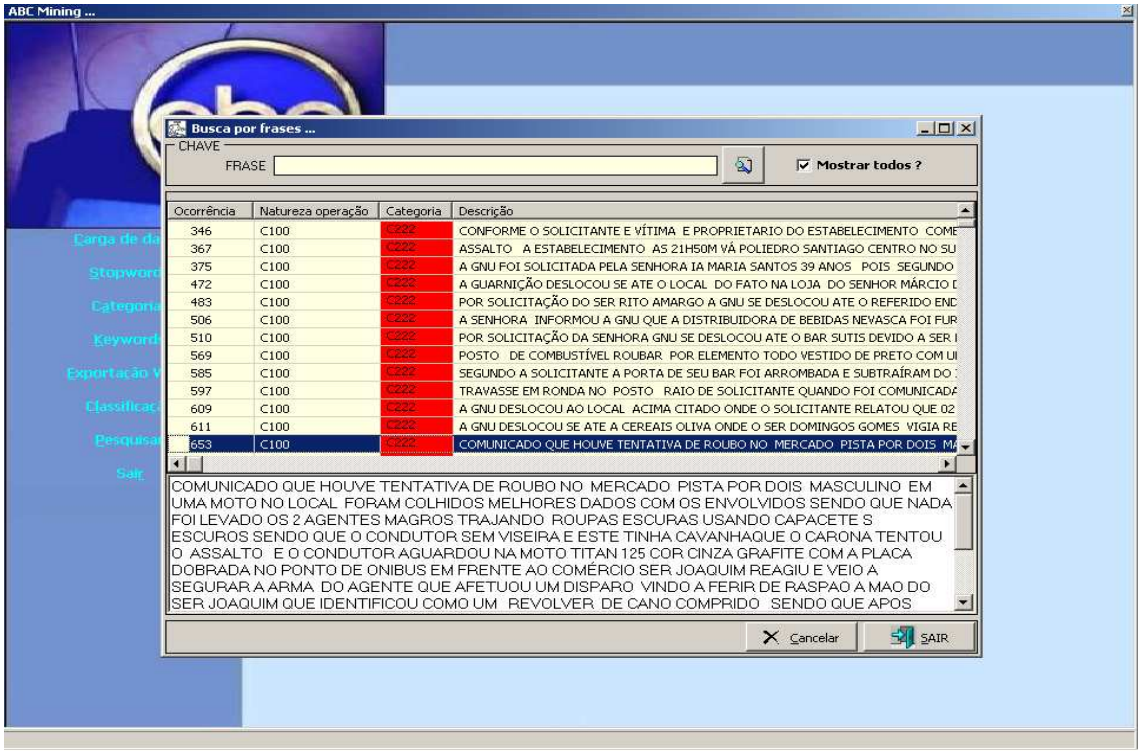


Figura 42 – Teste realizado para verificação do funcionamento do modelo gerado

4.3.1.1 Pesquisando na base de dados

Esta opção permite a pesquisa por *Keywords* ou por frases. O usuário pode selecionar a *Keyword* desejada visualizando todos os registros que contém essa palavra. A pesquisa por frase funciona da mesma maneira, só que permite buscar por um conjunto de palavras (Figura 43 e 44).

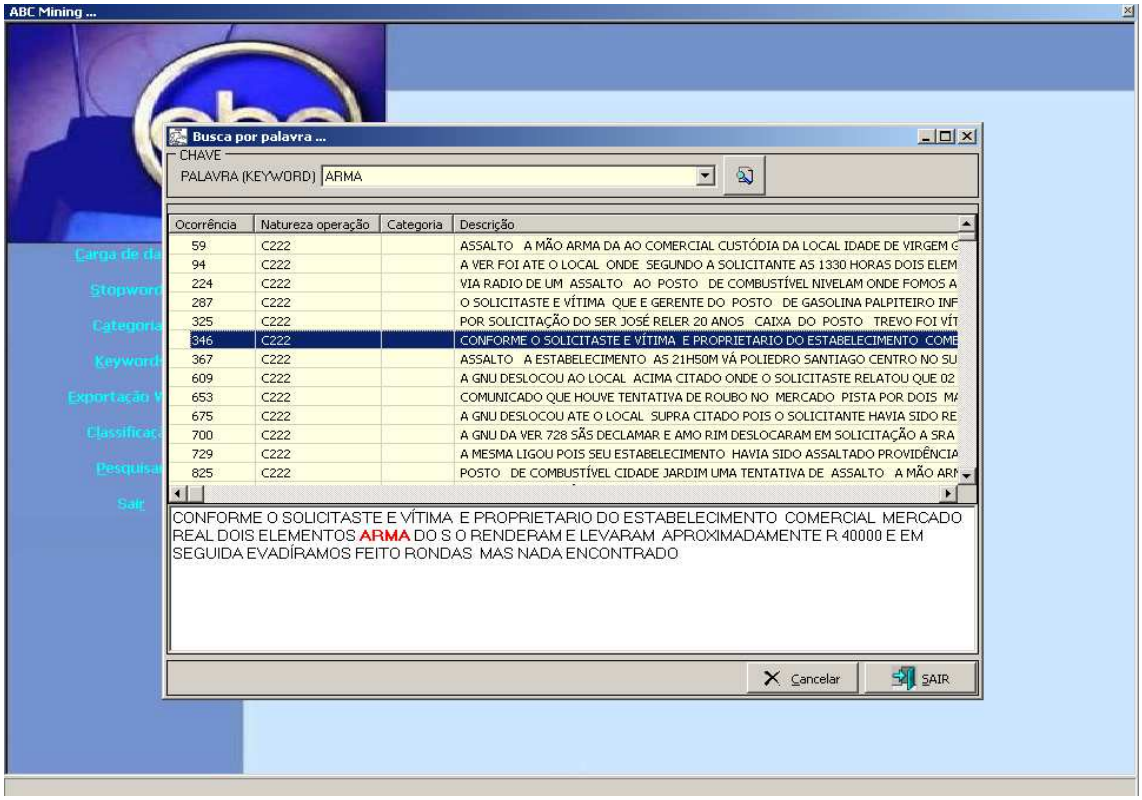


Figura 43 – Pesquisa pela *keyword* ARMA

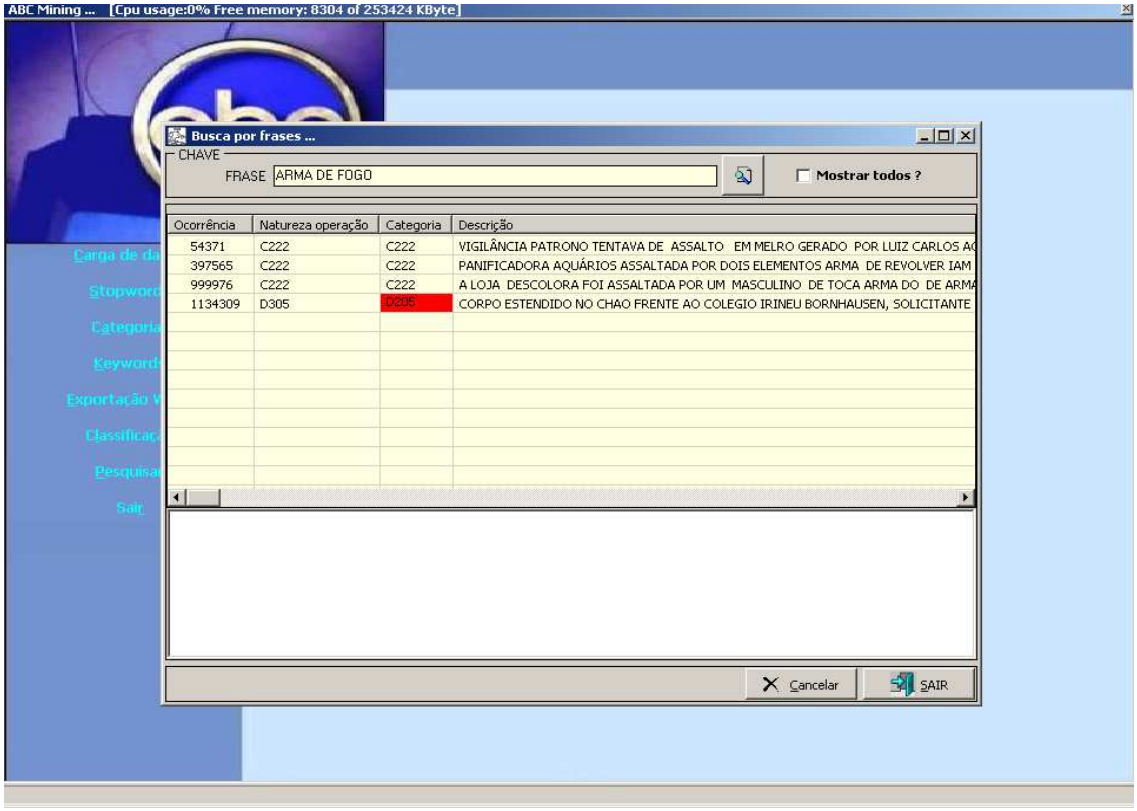


Figura 44 – Pesquisa por “Arma de fogo”

5 CONCLUSÕES E RECOMENDAÇÕES

Este trabalho propôs um modelo para reclassificação automática dos ROs, que permite a classificação automática do RO no momento da sua inclusão na base de dados da Secretaria de Segurança Pública e Defesa do Cidadão.

Os objetivos do trabalho foram alcançados. Como os dados estavam com muitos erros de ortografia, e com muitas palavras sem espaçamento correto, como, por exemplo, “armadefogo”, o *software* ABC *Clean* foi desenvolvido para solucionar esses problemas. O *software* atende as necessidades quanto a limpeza da base de dados e o agrupamento de palavras semelhantes. O ABC *Mining* atingiu o seu objetivo, conseguindo encontrar as *stopwords*, permitiu cadastrar as *keywords* enviadas pela DIRC e encontrar outras automaticamente, gerar o arquivo para que o *Weka*® pudesse gerar a árvore de decisão e reclassificar os ROs. O *Weka*® foi de grande ajuda para a geração da árvore de decisão. Mostrou-se eficiente na geração das regras. O ABC *Transform* surgiu durante o desenvolvimento dessa dissertação, da necessidade de transformar essa árvore de decisão em regras do tipo Se...Então para que o ABC *Mining* conseguisse entender as regras produzidas pelo *Weka*®.

Dos 2.684 ROs enviados pela DIRC, 5,4396% deles encontravam-se com a natureza de operação erradas (146 ROs). Esse percentual tende a aumentar, com a utilização de uma base de dados maior, e sem a seleção prévia dos possíveis registros corretos. Utilizou-se aproximadamente 50% dos ROs atendidos em 1 dia de atendimento do serviço 190. O modelo proposto mostrou-se eficaz para esses ROs e está apto a ser utilizado com outras bases de dados.

Quanto a tecnologia envolvida, acredita-se que está apenas na fase de divulgação e passará a ser aplicada em outras áreas, tais como: resposta automática no recebimento de e-mails, avaliando o conteúdo, auxiliando o *call center*; na classificação de notícias divulgadas pelos veículos de comunicação escrita; pode ajudar na avaliação da satisfação dos clientes (de forma qualitativa), etc...

Como sugestão para trabalhos futuros pode-se citar:

1) desenvolver um *software* capaz de dividir um registro de ocorrência em vários registros (exemplo: agressão seguida de assalto e morte). Neste

caso, o ideal seria 1 ROs com 3 delitos cadastrados. Para efeito estatístico muda de forma significativa os resultados, pois seriam computados 1 RO C115-Vias de fato ou agressão, 1 RO C221-Roubo ou assalto contra pessoa e outro RO C104- Homicídio;

2) estudar e aplicar técnicas de *cluster* nos ROs, tendo visto que hoje existem 732 diferentes naturezas de operação, porém essas não atendem completamente as necessidades de classificação. Como exemplo, tem-se a natureza de operação C222 – assalto a estabelecimento. Nessa natureza de operação estão classificados os assaltos a bares, padarias, lotéricas, lojas, etc... Em uma análise de *cluster*, novas naturezas de operação poderiam ser identificadas ou outras agrupadas em uma única. No exemplo acima, poderiam ser identificadas naturezas de operação específicas para cada tipo de estabelecimento;
3) o *software* ABC *Mining* poderia permitir o cadastro de novas regras automaticamente sem ter que alterar o código fonte do *software*.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AZEVEDO, Israel Belo de. **O prazer da produção científica:** diretrizes para a elaboração de trabalhos acadêmicos. 7.ed. Piracicaba: Ed. da Unimep, 1999. 208p.

BRASIL. Christiane Regina Soares. **Ferramenta inteligente de apoio à pesquisa:** mineração de artigos científicos na web. 2004. 62p. Monografia (Instituto de Ciências Matemáticas e de Computação) Universidade de São Paulo, São Carlos, 2004.

BRASIL. Ministério da Justiça. **Home institucional conceitos básicos**. Disponível em: <http://www.mj.gov.br/senasp/senasp/inst_conceitos.htm>. Acesso em: 05 abril, 2005.

BRAZDIL, Pavel. **Construção de modelos de decisão a partir de dados.** Disponível em <http://www.niaad.liacc.up.pt/~pbrazdil/Ensino/ML/DecTrees.html> Acesso em: 19 novembro, 2004.

CERQUEIRA, Daniel (Org.); LEMGRUBER, Julita (Org.); e MUSUMECI, Leonarda (Org.). Cadernos do Fórum de Debates sobre Criminalidade, Violência e Segurança Pública no Brasil. Rio de Janeiro : IPEA e CESeC/UCAM, 5 volumes, 2000.

CHAVES, Maurício Silveira. **Mapeamento e comparação de similaridade entre estruturas ontológicas.** 2004. 119f. Dissertação (Mestrado em Ciências da Computação) Pontífica Universidade Católica do Rio Grande do Sul, Porto Alegre, 2004.

CHEN, Hsinchun; CHUNG, Wingyan; XU. Jennifer Jie; WANG, Gang; QIN, Yi; CHAU, Michael. Crime data mining: a general framework and some examples. **IEEE Journal Computer**, v.37, 50-56, 2004.

CRISP-DM. **CRISP-DM:** Process Model.<http://www.crips-dm.org> Acesso em: 16 Fevereiro, 2005.

FAUSETT, Laurene V. **Fundamentals of neural networks**: architetures, algorithms, and applications. Englewood Cliffs. Prentice-Hall Inc, 1994. 461p.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social.** 5. ed. São Paulo: Atlas, 1999. 208p.

GNU. **The GNU Operation System**. Disponível em: <http://www.gnu.org>. Acesso em: 31 março, 2005.

HAN, Jiawei. KAMBER, Micheline. **Data Mining:** concepts and techniques. Morgan Kaufmann Publishers, New York, USA, 2001. 550p.

JESUS, Alberto Pereira de. **Data Mining aplicado à identificação do perfil dos usuários de uma biblioteca para a personalização de sistemas Web de recuperação e disseminação de informações.** 2004. 120f. Dissertação (Mestrado em Ciências da Computação) Universidade Federal de Santa Catarina, Florianópolis, 2004.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica.** 5.ed. São Paulo: Atlas, 2003. 312p.

LANDIS, Jr.; KOCH, GG. **The measurement of observer agreement for categorical data**. Biometrics, 1977; 33: 159-174.

LOH, Stanley; WIVES, Leandro Krug; OLIVEIRA, José Palazzo Moreira de. Descoberta proativa de conhecimento em coleções textuais: iniciando sem hipóteses. In: OFICINA DE INTELIGÊNCIA ARTIFICIAL, 4. 2000. Pelotas. **Proceedings...** Pelotas: EDUCAT, 2000. 143-154p.

LOH, Stanley; OLIVEIRA, José Palazzo Moreira de; GAMEIRO, Maurício Almeida. Knowledge discovery in texts for constructing decision support systems. **The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies - Special Issue on Text and Web Mining,** Journal of Applied Intelligence, Kluwer Academic Publishers, v.18, p.357-366, 2003. ISSN: 0924-669X.

LOH, Stanley. **Data mining**. Disponível em: <http://atlas.ucpel.tche.br/~loh/dm-ppt.pdf>. Acesso em: 30 de maio, 2005.

LOUZADA NETO, Francisco. DINIZ, Carlos Alberto R. **Data mining:** uma introdução. São Paulo. Associação Brasileira de Estatística, 2000. 123p.

MANNING, Christopher D.; SHÜTZE, Hinrich. **Foundations of statistical natural language processing.** Cambridge. The Mit Press, 1999. 620p.

MATTAR, Fauze Najib. **Pesquisa de marketing**. São Paulo: Atlas, 1999. 344p.

MUNIZ, Jacqueline de Oliveira. **Registros de ocorrência da PCERJ como fonte de informações criminais.** In: Criminalidade, violência e segurança pública no brasil: uma discussão sobre as bases de dados e questões metodologicas, 2000, Rio de Janeiro. Fórum de debates CESeC/ IPEA. Rio de Janeiro: IPEA e Centro de Estudos de Segurança e Cidadania - CESESC, 2000.

PORTER, Martin F. **An algorithm for suffix stripping**. In Readings in Information Retrieval, 313-316. Morgan Kaufmann, 1997.

PRADO, Hércules Antonio do. **Conceitos de descoberta de conhecimento em bancos de dados.** 1997. 43f. Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1997.

REZENDE. Solange Oliveira. **Sistemas inteligentes:** fundamentos e aplicações. São Paulo: Manole, 2003. 525p.

ROESCH, Sylvia Maria Azevedo. **Projetos de estágio e de pesquisa em administração:** guia para estágios, trabalhos de conclusão, dissertações e estudos de caso. São Paulo: Altas, 1999. 304p.

SANTA CATARINA. Polícia Militar de Santa Catarina. **Um pouco de história.** Disponível em:<http://www.pm.sc.gov.br/website/redir.php?site=40&act=1&id=4>. Acesso em: 15 abril, 2004.

SANTOS, Maria Angela Moscalewski Roveredo dos. **Extraindo regras de associação a partir de textos.** 2002. 51f. Dissertação (Mestrado em Informática Aplicada) – Pontífica Universidade Católica do Paraná, Curitiba, 2002.

SARDINHA, Tony Berber. **O banco de palavras-chave.** Disponível em: < http://sites.uol.com.br/tony4/homepage.html>. Acesso em: 23 março. 2004.

SILVA, Edilberto Magalhães. **Descoberta de conhecimento com o uso de text mining:** cruzando o abismo de Moore**.** 2002. 175f. Dissertação de Mestrado em Gestão do Conhecimento e Tecnologia da Informação, Universidade Católica de Brasília, Brasília, 2002.

SSPSC. Secretaria de Segurança Pública e Defesa do Cidadão. **Home SSP**. Disponível em: <http://www.ssp.sc.gov.br/dirc/historico.htm>. Acesso em: 10 novembro, 2004.

TAN, Ah-Hwee. **Text Mining:** the state of the art and the challenges. Kent Ridge Digital Labs, 1999. Disponível em: <http://textmining.krdl.org.sg>. Acesso em: 19 setembro. 2003.

WEKA. The University Of Waikato. **Data mining with open source machine learning software in Java**. Disponível em: <www.cs.waikato.ac.nz/ ml/weka>. Acesso em: 2 fevereiro, 2005.

WIVES, Leandro Krug. **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de *Clustering***. Dissertação (Mestrado em Ciência da Computação). Programa de Pós-graduação em Gestão do conhecimento da Universidade Federal do Rio Grande do Sul, Porto Alegre, 1999, 84p.

WIVES, Leandro Krug. Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência Competitiva. Disponível em: <http://www.inf.ufrgs.br/~wives/portugues/publicacoes.html>. Acesso em: 19 setembro, 2003.

YANG, Yiming; LIU, Xin. An evaluation of statistical approaches to text-categorization. **Journal of Information Retrieval**, Kluwer Academic Publishers, v.1, n.1/2, p.69-90. 1999.

ANEXO A – TABELA COM A DESCRIÇÃO DAS NATUREZAS DE OPERAÇÃO

Existem 732 códigos de natureza de operação, sendo que nessa dissertação foram estudados 17, que estão descritos na Tabela 7 a seguir:

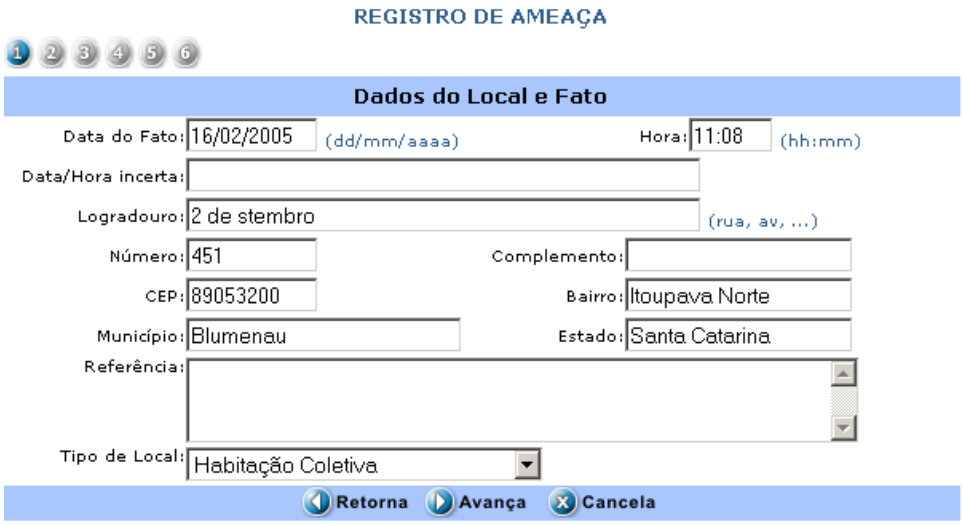
Tabela 7 - Naturezas de operação estudadas

Natureza de Operação	Descrição
A203	apoio/reforço à polícia civil
C104	homicídio
C112	seqüestro e/ou cárcere privado
C114	tentativa de homicídio
C207	extorsão mediante seqüêstro
C215	roubo consumado
C221	roubo ou assalto contra pessoa
C222	roubo ou assalto a estabelecimento
C610	disparo de arma de fogo
C903	comunicação falsa
D205	encontro de cadáver
D305	endereço incompleto
D309	óbito no local
D316	vítima já conduzida por populares
E111	ferimento por arma branca
E112	ferimento por arma de fogo
E123	traumatismo crâneo-encefálico

ANEXO B – ETAPAS DO REGISTRO DE OCORRÊNCIA PELA INTERNET

Este sistema foi desenvolvido pelo Centro de Informática e Automação do Estado de Santa Catarina (CIASC), seu objetivo é permitir ao cidadão efetuar registros e consultas de alguns tipos de ocorrências policiais previamente liberadas pela SSP na internet. Neste exemplo fez-se a inclusão de um registro de ameaça. Esse registro é composto de 6 etapas: dados do local e fato, dados pessoais do comunicante, endereço do comunicante, relação de documentos, dados do autor, histórico do fato ocorrido.

Etapa 1: O ameaçado deve informar dados do local aonde ele está sendo ameaçado e a data e hora que ocorreu ou está ocorrendo.

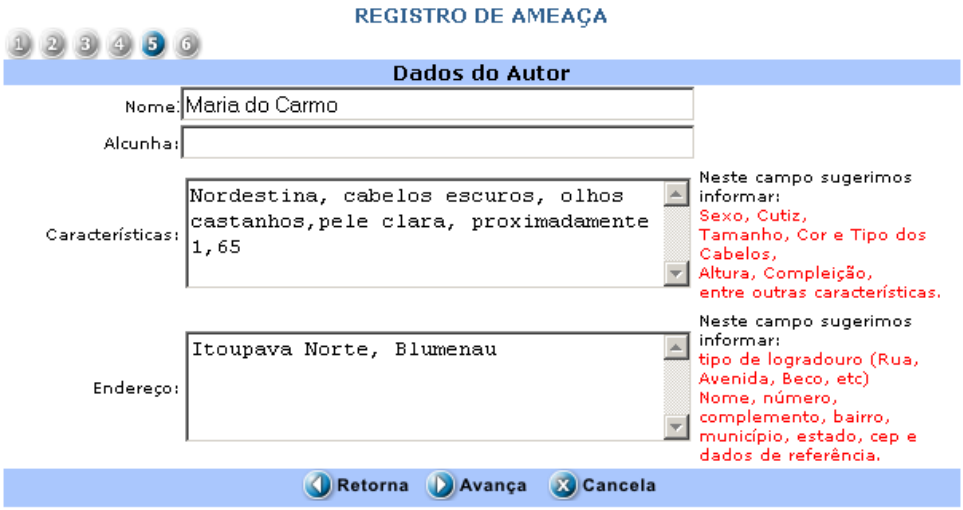


Etapa 2: Documentos pessoais, local de nascimento e filiação devem ser preenchidos nessa etapa.

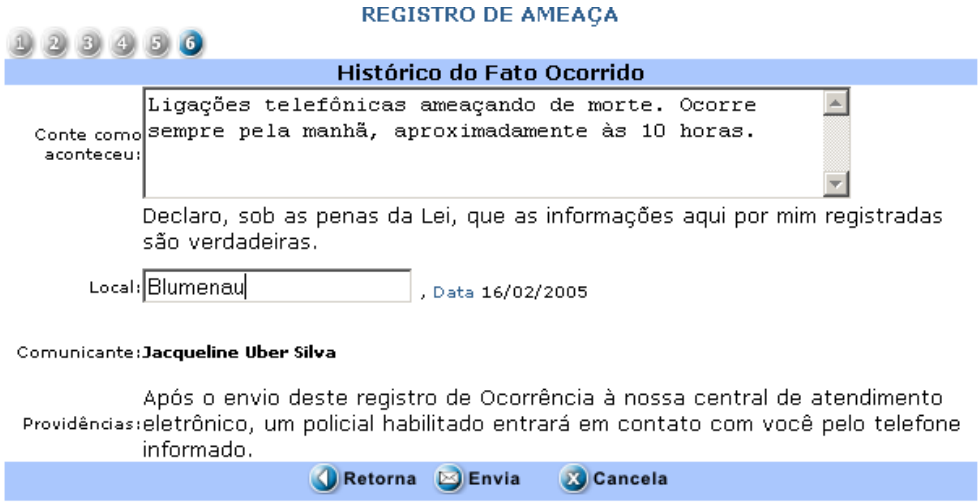
Etapa 3: Como o registro de ocorrência pode ser efetuado por outra pessoa sem ser o próprio ameaçado, nessa etapa os dados do endereço do comunicante devem ser preenchidos para posterior confirmação pela Polícia Militar.

Etapa 4: essa etapa somente é preenchida em registro de perda de documentos, não sendo requerida no registro de ameaça, pois serve para relacionar os documentos perdidos.

Etapa 5: dados da pessoa que está fazendo as ameaças devem ser informadas aqui. As características do ameaçador devem ser informadas para auxiliar a Polícia Militar na sua identificação.



Etapa 6: uma narração de como aconteceu a ameaça deve ser informada e o local aonde aconteceu. Assim que for confirmado o envio desse registro, um policial entrará em contato por telefone com o comunicante.



ANEXO C – BASE DE TREINAMENTO

ocor	natur	historico
62	C222	informou que houve um assalto na panificadora da´magro as gus pm 12
94	C222	a vtr foi ate o local onde segundo a solicitante as 1330 horas dois
346	C222	conforne o solicitante e vitima e proprietario do estabelecimento c
367	C222	assalto a estabelecimento as 21h50m av polidoro santiago centro no
375	C222	a gu foi solicitada pela sra iza maria santos 39 anos pois segundo
506	C222	a sra neli informou a gu que a distribuidora de bebidas nevasca foi
609	C222	a gu deslocou ao local acima citado onde o solicitante relatou que 0
653	C222	comunicado que houve tentativa de roubo no mercado pisetta por dois
700	C222	a gu da vtr 728 sds laudeci vilmar e amorim deslocaram em solicitaca
729	C222	a mesma ligou pois seu estabelecimento havia sido assaltado providê
793	C222	a gu deslocou se ate o local acima citado por volta das 1900 dois m
1039	C222	segundo a solicitante dois elementos armados de revolveres assaltar
1234	C222	a gu da vtr 1926 composta pelos pms sd amorim cardoso nascimento e
1339	C222	por sol da 4 cia as vtrs foram no local onde conseguiram intercepta
1436	C222	a vtr com a gu encontrava se em rondas no centro quando recebeu a c
1506	C222	segundo o solicitante dois elementos com revolveres calibres 38 co
4388	C222	o copom foi informado que dois elementos armados roubaram do bar 3
5064	C222	masculinos encapussados armados de faca assaltam clube aimore gerad
5964	C222	o solicitante comunicou o copom que tinham arrombado o vidro da por
6074	C222	dois masculinos com veiculo moto onde um destes moreno estatura al
6704	C222	assalto a estabelecimento comercial bingo gerado por claudio jus
8518	C222	gerado por altair oliverio da costa as 1204 hs 2222 encerramento
10056	C222	o solicitante proprietario da danceteria balanga tanga informou qu
11910	C222	roubo assalto em estabelecimento comercial gerado por joao mari
12180	C222	esmo imforma que jovem arrombou trabucao gerado por copom as 0429
12193	C222	entativa de assalto no mercado sasse gerado por andreia geovan
12276	C222	solic inf roubo em estabelecimento comercial gerado por marcelo
12745	C222	informa que foi assaltada gerado por adayir freitas bittencour as

12746	C222	informa que foi assaltada a farmacia litoral gerado por adayir fr
12759	C222	padaria sao jose masculino armado assalto estabelecimento gera
12909	C222	solicitante informa que seu estabelecimento foi assaltado gerado
13378	C222	informa assalto a mao armada por dois masculinos armados gerado
1048771	E112	masculino baleado no pe encontrado pela gu pm proximo da inte
1049986	C221	embaixo da ponte masculino foi abordado no semafaro da beira ma
1052749	D309	na rua do rodeio perto do campo de futebol masculino alvejado
1052964	C104	proximo conjunto habitacional dona blides masculino baleado n
1053084	D309	esquina cristal de araujo tiroteio jaqueta preta com uma listra
1053561	C222	peche novo atraz da pedrita na rua otavio cruz n 451feminina in
1053590	C207	condutora de veiculo foi vitima de sequestro relampago gerado
1058663	C104	proximo ao bar do benedito femininainino sem sinais vitais ge
1058693	D309	em frente a casan masculino caido na via com hemorragia pelo n
1058747	D309	na subida da rua dois elementos alvejados estao caidos nos fund
1059862	D309	na favela disparos de arma de fogo no local gerado por valcel
1062707	C104	no bairro zanelato no final da rua beija flor disparo de arma
1063897	C114	altamiro guimaraes na frente do imperatriz masculino balead
1064191	D205	proximo bar do zica na casa da dona catarina de cor roza no be
1064433	C104	elemento morto no local homicidio na rua aguardando vtr gerado
1064740	D309	fundos da transol proximo mercado coelho masculino de bicic
1065685	C104	sujeito auvejado por arma de fogo disparada por dois sujeitos pi
1066745	C104	ten comunica que conforme informacao de um cidadao de nome alexa
1070007	D309	elemento vitima de tiro proximo do campo gerado por helio gi
1070406	D309	gerado por pedro paulo scremin marti as 14 12 hs 6025 encer
1072276	C215	no bairro bom abrigo veiculo roubado ha instantes mey 8630 cl
1073450	E112	no bar do gordo proximo ao casa nova masculino de jaqueta a
1075384	D316	proximo posto de saude masculino baleado gerado por marcio
1076660	D309	masculino s passaram correndo na frente da residencia do solicit
1077051	C104	masculino baleado em residencia proximo agua pura praca paulo
1078178	C104	gerado por marco roberto frederico as 18 34 hs 6033 encerra
1080547	C112	carlos pol civil comunica que duas feminina foram tomadas de re

1084035	C104	rua antes do tunel de capoeiras masculino alvejado por arma de
1084130	C222	apos a igreja assembleia de deus mercearia ani foiassaltada a ce
1084898	C221	depois do recanto das pedras furto em veiculo gol vermelho depo
1087264	E123	masculino foi agredido com um extinto na cabeca esta inconscien
1089775	E112	ao lado do banco do besc na farmacia zachi motoqueiro levou um t
1090200	E112	edificio esteves junior casal brigando no interior do edificio
1090623	E111	serv maria helena masculino alvejado gerado por rosana ramo
1091030	C104	comunica que chegou masculino baleado no hrsj de nome fabricio
1091296	C112	sd belmiro de itapema comunica que um casal foi vitima de seques
1091298	C112	proximo ao tunel no posto galo masc vitima de sequestro relamp
1092420	C104	masculino morto na marginal da 101 ao lado do ctg gerado por
1093950	D309	frente ao bar do tio ze masculino baleado com tres tiros gera
1094610	D309	lado da igreja universal segundo informacoes masculino foi a
1094646	C104	proximo ao bar cal masculino baleado gerado por alan assunc
1094948	D205	em frente ao mercado alemao em um pasto masculino encontrado p
1095970	D309	disparo de arma de fogo no final desta rua acima citada e uso de
1097573	C221	roubo contra pessoa no caixa eletronico do bradesco esta ocorre
1098341	C104	proximo campo do nacional 3 pessoas feridas com arma de fogo so
1098643	D205	proximo a panificadora emilio encontraram um corpo gerado po
1099399	D309	veiculo chocou se contra um poste nos fundos da antiga prefeitur
1100022	D205	masculino em obito proximo mercearia da eliane rua da fonte
1100928	C104	proximo ao lanchonete daniela masculino baleado no local gera
1101855	C221	sol foi tomado em assalto com seu veiculo pagero preta nova mb
1102085	C221	proximo a panificadora leal feminina foi assaltada gerado po
1102468	D309	antes de chegar nos ingleses proximo do texaco perda da casa n
1106374	C610	proximo colegio pero vaz sol relata que esta ouvindo disparo d
1106721	E112	elemento alvejado em frente cantuaria gerado por etevaldo fer
1106934	C610	disparo de arma de fogo por 03 masculino s um de camiseta cre
1107211	C221	vitima de assalto gerado por luiz fernando cardoso as 22 23
1107521	C221	furtado proximo a boat x fiat uno mille fire cor cinza placa
1108509	E111	masculino armado de revolver ameacando o pai do solic defront

1108539	C221	assalta a mao armada contra pessoa no patio do koxixo s bar ger
1108871	A203	apos hotel aguas mornas em um terreno baldio apoio a policia ci
1109227	C221	moto titan maj3990 verde 97 jose em direcao ao rio tavares moto
1111133	C221	proximo ao edificio deluno masculino informa que furtaram o m
1112914	C114	no real parque na rua domingos jose dos santos pro x a lagoa do
1117324	C221	agencia da caixa economica caixa eletronico masculino de bone
1118137	E112	jardim janaina masculino ferido por disparo de arma de fogo ag
1119840	D309	proximidades da igreja assembleia de deus masculino alvejado
1120303	C104	Disparo de armas de fogo no final da escadaria no morro da penit
1120325	C104	final da rua disparos de arma de fogo no local gerado por ma
1121326	D309	frente supermercado s sul do rio masculino foi esfaqueado num
1121950	C221	proximo fazenda professor nader masculino foi assaltado por va
1122100	C222	sol informa assalto na farm santo estevao meliantes deslocaram
1122139	C114	Disparo de arma de fogo proximo a kidoce gerado por gilson jo
1122368	C104	homicidio no posto de saude da tapera masculino teve um corte n
1122682	C221	Tomado de assalto por um elemento armado o veiculo renaul clio
1123313	E112	proximo a academia atlas masculino baleado no local gerado
1126479	C114	Disparo de arma de fogo masculino atingido com dois disparos n
1127609	C104	defronte a empresa macedo disparo de armas de fogo e um veiculo
1127744	D309	no final da rua entra a esquerda proximo da escadaria 05 mascu
1129512	E112	dois masculino baleados no local proximo ao trevo da seta ge
1129725	D316	proximo ao beira rio bar masculino alvejado na barriga gerad
1134309	D305	corpo estendido no chao frente ao colegio irineu bornhausen sol
1134704	C104	no morro do mocota na cabeca do santos tiroteio com vitimas g
1135031	C114	final da rua masculino foi agredido no bar do tico solicitam vt
1135220	E112	na entrada da estiva apos trevo de governador 1 entrada a esquer
1136964	D309	no catarina proximo madereira do osni masculino foi alvejado
1139904	D309	no jardim das laranjeiras na antiga fabrica de sabao residenci
1140717	C104	solicitante informa que no saco dos limoes masculino de alcunha
1141347	D305	no morro do flamengo proximo ao mini mercado valdir rammes ma
1141622	E112	disparo de arma de fogo na desida da praia proximo ao condominio

APÊNDICE A – Cálculos realizados para encontrar a entropia e ganho com a *keyword* vital

Para Vital = 1

$= - [0/30 \log_2 0/30 + 5/30 \log_2 5/30 + 0/30 \log_2 0/30 + 1/30 \log_2 1/30 + 0/30 \log_2 0/30 + 0/30 \log_2 0/30 + 0/30 \log_2 0/30 + 0/30 \log_2 0/30 + 0/30 \log_2 0/30 + 1/30 \log_2 1/30 + 1/30 \log_2 1/30 + 2/30 \log_2 2/30 + 19/30 \log_2 19/30 + 0/30 \log_2 0/30 + 0/30 \log_2 0/30 + 1/30 \log_2 1/30 + 0/30 \log_2 0/30]$

I = 1,7628818

Para Vital = 0

$= - [1/30 \log_2 1/30 + 15/30 \log_2 15/30 + 3/30 \log_2 3/30 + 4/30 \log_2 4/30 + 1/30 \log_2 1/30 + 1/30 \log_2 1/30 + 13/30 \log_2 13/30 + 35/30 \log_2 35/30 + 1/30 \log_2 1/30 + 3/30 \log_2 3/30 + 0/30 \log_2 0/30 + 0/30 \log_2 0/30 + 2/30 \log_2 2/30 + 2/30 \log_2 2/30 + 9/30 \log_2 9/30 + 1/30 \log_2 1/30]$

I = 2,8128

E(Vital) = 30/121 . 1,76288 + 91/121 . 2,8128

E(Vital) = 2,5522

Gain (Vital) = 3,114205 – 2,5522 = 0,561981

APÊNDICE B – Regras geradas pelo *Weka*®

Partindo-se da base do arquivo ocorrencia.arff gerado pelo *software* ABC *Mining* o *software Weka*® gerou a seguinte árvore de decisão:

VITAIS = 1
| BAIRRO = 1
| | SANGUE = 1
| | | CIVIL = 1: D205
| | | CIVIL = 0: D305
| | SANGUE = 0
| | | ALVEJADO = 1
| | | | CORPO = 1: C903
| | | | CORPO = 0: C104
| | | ALVEJADO = 0: C222
| BAIRRO = 0
| | ASSALTO = 1
| | | ALVEJADO = 1: D309
| | | ALVEJADO = 0: C222
| | ASSALTO = 0
| | | CORPO = 1
| | | | AGENTE = 1
| | | | | DISPARO = 1: C104
| | | | | DISPARO = 0: E112
| | | | AGENTE = 0
| | | | | ALVEJADO = 1
| | | | | | DISPARO = 1: D309
| | | | | | DISPARO = 0: C104
| | | | | ALVEJADO = 0: C114
| | | CORPO = 0
| | | | CIVIL = 1

| | | | | ALVEJADO = 1: D309
| | | | | ALVEJADO = 0
| | | | | | POPULARES = 1: D309
| | | | | | POPULARES = 0: C610
| | | | CIVIL = 0
| | | | | MOTO = 1: C104
| | | | | MOTO = 0
| | | | | | FOGO = 1
| | | | | | | GUARNICAO = 1: D309
| | | | | | | GUARNICAO = 0
| | | | | | | | AGENTE = 1: D309
| | | | | | | | AGENTE = 0
| | | | | | | | | ALVEJADO = 1
| | | | | | | | | | POPULARES = 1
| | | | | | | | | | | DISPARO = 1: C104
| | | | | | | | | | | DISPARO = 0: D309
| | | | | | | | | | POPULARES = 0: D309
| | | | | | | | | ALVEJADO = 0
| | | | | | | | | | POPULARES = 1: D309
| | | | | | | | | | POPULARES = 0: C104
| | | | | | FOGO = 0: D309
VITAIS = 0
| ASSALTO = 1
| | TOMADO = 1
| | | HOSPITAL = 1
| | | | ARMA = 1: C222
| | | | ARMA = 0: C112
| | | HOSPITAL = 0
| | | | ESTABELECIMENTO = 1: C222
| | | | ESTABELECIMENTO = 0
| | | | | POSTO = 1
| | | | | | AGENTE = 1: C222

| | | | | | AGENTE = 0
| | | | | | | FEMININA = 1
| | | | | | | | ARMA = 1: C222
| | | | | | | | ARMA = 0: C222
| | | | | | | FEMININA = 0: C222
| | | | | POSTO = 0
| | | | | | CIVIL = 1: C222
| | | | | | CIVIL = 0
| | | | | | | FOGO = 1: C222
| | | | | | | FOGO = 0
| | | | | | | | BAIRRO = 1: C222
| | | | | | | | BAIRRO = 0
| | | | | | | | | MOTO = 1: C222
| | | | | | | | | MOTO = 0
| | | | | | | | | | AGENTE = 1
| | | | | | | | | | | FEMININA = 1: C221
| | | | | | | | | | | FEMININA = 0
| | | | | | | | | | | | ARMA = 1: C222
| | | | | | | | | | | | ARMA = 0: C222
| | | | | | | | | | AGENTE = 0
| | | | | | | | | | | ARMA = 1
| | | | | | | | | | | | FEMININA = 1: C222
| | | | | | | | | | | | FEMININA = 0: C221
| | | | | | | | | | | ARMA = 0: C222
| | TOMADO = 0
| | | FEMININA = 1
| | | | RESIDENCIA = 1
| | | | | AGENTE = 1: C222
| | | | | AGENTE = 0: C112
| | | | RESIDENCIA = 0
| | | | | CIVIL = 1
| | | | | | ESTABELECIMENTO = 1: C222

| | | | | | ESTABELECIMENTO = 0: C112
| | | | | CIVIL = 0
| | | | | | MORRO = 1: C221
| | | | | | MORRO = 0
| | | | | | | HOSPITAL = 1
| | | | | | | | POSTO = 1: C222
| | | | | | | | POSTO = 0: C221
| | | | | | | HOSPITAL = 0
| | | | | | | | FOGO = 1
| | | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | | ESTABELECIMENTO = 0
| | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | POSTO = 0: C222
| | | | | | | | FOGO = 0
| | | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | | ESTABELECIMENTO = 0
| | | | | | | | | | MOTO = 1: C222
| | | | | | | | | | MOTO = 0
| | | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | | POSTO = 0
| | | | | | | | | | | AGENTE = 1
| | | | | | | | | | | | ARMA = 1: C222
| | | | | | | | | | | | ARMA = 0: C222
| | | | | | | | | | | AGENTE = 0
| | | | | | | | | | | | ARMA = 1: C222
| | | | | | | | | | | | ARMA = 0
| | | | | | | | | | | | | BAIRRO = 1: C222
| | | | | | | | | | | | | BAIRRO = 0
| | | | | | | | | | | | | | CORPO = 1: C222
| | | | | | | | | | | | | | CORPO = 0
| | | | | | | | | | | | | | | HOTEL = 1: C222
| | | | | | | | | | | | | | | HOTEL = 0: C222

| | | FEMININA = 0
| | | | ESTABELECIMENTO = 1: C222
| | | | ESTABELECIMENTO = 0
| | | | | FOGO = 1
| | | | | | AGENTE = 1
| | | | | | | DISPARO = 1
| | | | | | | | BAIRRO = 1: C222
| | | | | | | | BAIRRO = 0
| | | | | | | | | POSTO = 1: C222
| | | | | | | | | POSTO = 0
| | | | | | | | | | MOTO = 1: C222
| | | | | | | | | | MOTO = 0: C222
| | | | | | | DISPARO = 0: C222
| | | | | | AGENTE = 0: C222
| | | | | FOGO = 0
| | | | | | ARMA = 1
| | | | | | | MOTO = 1: C222
| | | | | | | MOTO = 0
| | | | | | | | RESIDENCIA = 1
| | | | | | | | | POSTO = 1: C221
| | | | | | | | | POSTO = 0: C222
| | | | | | | | RESIDENCIA = 0
| | | | | | | | | BAIRRO = 1
| | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | POSTO = 0
| | | | | | | | | | | AGENTE = 1: C222
| | | | | | | | | | | AGENTE = 0
| | | | | | | | | | | | MORRO = 1: C222
| | | | | | | | | | | | MORRO = 0
| | | | | | | | | | | | | CIVIL = 1: C222
| | | | | | | | | | | | | CIVIL = 0: C222
| | | | | | | | | BAIRRO = 0

| | | | | | | | | CIVIL = 1
| | | | | | | | | | POSTO = 1
| | | | | | | | | | AGREDIDO = 1: C222
| | | | | | | | | | AGREDIDO = 0: C222
| | | | | | | | | | POSTO = 0: C222
| | | | | | | | | CIVIL = 0
| | | | | | | | | | HOTEL = 1
| | | | | | | | | | | AGENTE = 1
| | | | | | | | | | | | GUARNICAO = 1: C222
| | | | | | | | | | | | GUARNICAO = 0: C222
| | | | | | | | | | | AGENTE = 0: C222
| | | | | | | | | | HOTEL = 0
| | | | | | | | | | | GUARNICAO = 1
| | | | | | | | | | | | POSTO = 1
| | | | | | | | | | | | | FATAL = 1: C222
| | | | | | | | | | | | | FATAL = 0: C221
| | | | | | | | | | | | POSTO = 0: C222
| | | | | | | | | | | GUARNICAO = 0
| | | | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | | | POSTO = 0
| | | | | | | | | | | | | AGENTE = 1
| | | | | | | | | | | | | | DISPARO = 1: C222
| | | | | | | | | | | | | | DISPARO = 0
| | | | | | | | | | | | | | | POPULARES = 1: C222
| | | | | | | | | | | | | | | POPULARES = 0
| | | | | | | | | | | | | | | | HOSPITAL = 1: C222
| | | | | | | | | | | | | | | | HOSPITAL = 0
| | | | | | | | | | | | | | | | | SANGUE = 1: C222
| | | | | | | | | | | | | | | | | SANGUE = 0
| | | | | | | | | | | | | | | | | | AGREDIDO = 1: C222
| | | | | | | | | | | | | | | | | | AGREDIDO = 0
| | | | | | | | | | | | | | | | | | | CORPO = 1: C222

| | | | | | | | | | | | | | | | | | | | | CORPO = 0
| | | | | | | | | | | | | | | | | | | | | | MORRO = 1: C222
| | | | | | | | | | | | | | | | | | | | | | MORRO = 0: C222
| | | | | | | | | | | | | | | | AGENTE = 0
| | | | | | | | | | | | | | | | DISPARO = 1: C222
| | | | | | | | | | | | | | | | DISPARO = 0
| | | | | | | | | | | | | | | | MORRO = 1: C222
| | | | | | | | | | | | | | | | MORRO = 0
| | | | | | | | | | | | | | | | | POPULARES = 1: C222
| | | | | | | | | | | | | | | | | POPULARES = 0
| | | | | | | | | | | | | | | | | | MANTEVE = 1: C222
| | | | | | | | | | | | | | | | | | MANTEVE = 0
| | | | | | | | | | | | | | | | | | | AGREDIDO = 1: C222
| | | | | | | | | | | | | | | | | | | AGREDIDO = 0
| | | | | | | | | | | | | | | | | | | | HOSPITAL = 1: C222
| | | | | | | | | | | | | | | | | | | | HOSPITAL = 0
| | | | | | | | | | | | | | | | | | | | | ALVEJADO = 1: C222
| | | | | | | | | | | | | | | | | | | | | ALVEJADO = 0
| | | | | | | | | | | | | | | | | | | | | | SANGUE = 1: C222
| | | | | | | | | | | | | | | | | | | | | | SANGUE = 0: C222
| | | | | ARMA = 0
| | | | | | AGENTE = 1: C222
| | | | | | AGENTE = 0
| | | | | | | MOTO = 1
| | | | | | | | POSTO = 1: C222
| | | | | | | | POSTO = 0
| | | | | | | | | BAIRRO = 1: C222
| | | | | | | | | BAIRRO = 0
| | | | | | | | | | DISPARO = 1: C222
| | | | | | | | | | DISPARO = 0
| | | | | | | | | | | GUARNICAO = 1: C222
| | | | | | | | | | | GUARNICAO = 0

| | | | | | | | | | | | | HOSPITAL = 1: C222
| | | | | | | | | | | | | HOSPITAL = 0
| | | | | | | | | | | | | | RESIDENCIA = 1: C222
| | | | | | | | | | | | | | RESIDENCIA = 0: C222
| | | | | | | | MOTO = 0
| | | | | | | | | POSTO = 1
| | | | | | | | | | BAIRRO = 1: C222
| | | | | | | | | | BAIRRO = 0
| | | | | | | | | | DISPARO = 1: C222
| | | | | | | | | | DISPARO = 0
| | | | | | | | | | | MORRO = 1: C222
| | | | | | | | | | | MORRO = 0
| | | | | | | | | | | | RESIDENCIA = 1: C222
| | | | | | | | | | | | RESIDENCIA = 0: C222
| | | | | | | | | POSTO = 0
| | | | | | | | | | AGREDIDO = 1: C222
| | | | | | | | | | AGREDIDO = 0
| | | | | | | | | | | DISPARO = 1: C222
| | | | | | | | | | | DISPARO = 0
| | | | | | | | | | | | GUARNICAO = 1: C222
| | | | | | | | | | | | GUARNICAO = 0
| | | | | | | | | | | | | BAIRRO = 1: C222
| | | | | | | | | | | | | BAIRRO = 0
| | | | | | | | | | | | | | HOSPITAL = 1: C222
| | | | | | | | | | | | | | HOSPITAL = 0
| | | | | | | | | | | | | | | HOTEL = 1: C222
| | | | | | | | | | | | | | | HOTEL = 0
| | | | | | | | | | | | | | | | POPULARES = 1: C222
| | | | | | | | | | | | | | | | POPULARES = 0: C222
| ASSALTO = 0
| | DISPARO = 1
| | | POPULARES = 1

| | | | ESTABELECIMENTO = 1: C222
| | | | ESTABELECIMENTO = 0
| | | | | BAIRRO = 1: D316
| | | | | BAIRRO = 0
| | | | | | ARMA = 1: C104
| | | | | | ARMA = 0
| | | | | | | ALVEJADO = 1
| | | | | | | | HOSPITAL = 1: C104
| | | | | | | | HOSPITAL = 0: D316
| | | | | | | ALVEJADO = 0: E112
| | | POPULARES = 0
| | | | FEMININA = 1
| | | | | AGENTE = 1
| | | | | | ALVEJADO = 1: E112
| | | | | | ALVEJADO = 0: E111
| | | | | AGENTE = 0
| | | | | | CIVIL = 1: C222
| | | | | | CIVIL = 0: C104
| | | | FEMININA = 0
| | | | | FOGO = 1
| | | | | | HOSPITAL = 1
| | | | | | | AGENTE = 1: E112
| | | | | | | AGENTE = 0
| | | | | | | | CIVIL = 1: C222
| | | | | | | | CIVIL = 0
| | | | | | | | | MOTO = 1: C114
| | | | | | | | | MOTO = 0: C114
| | | | | | HOSPITAL = 0
| | | | | | | CORPO = 1: C104
| | | | | | | CORPO = 0
| | | | | | | | ALVEJADO = 1: E112
| | | | | | | | ALVEJADO = 0

| | | | | | | | | TOMADO = 1: C104
| | | | | | | | | TOMADO = 0
| | | | | | | | | | BAIRRO = 1
| | | | | | | | | | | MOTO = 1: C222
| | | | | | | | | | | MOTO = 0: C104
| | | | | | | | | | BAIRRO = 0: C222
| | | | | FOGO = 0
| | | | | | RESIDENCIA = 1: C104
| | | | | | RESIDENCIA = 0
| | | | | | | ARMA = 1: C222
| | | | | | | ARMA = 0
| | | | | | | | AGENTE = 1: C222
| | | | | | | | AGENTE = 0
| | | | | | | | | ALVEJADO = 1: C222
| | | | | | | | | ALVEJADO = 0
| | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | POSTO = 0: C114
| | DISPARO = 0
| | | CORPO = 1
| | | | FOGO = 1: D205
| | | | FOGO = 0
| | | | | CIVIL = 1
| | | | | | ESTABELECIMENTO = 1: C104
| | | | | | ESTABELECIMENTO = 0: A203
| | | | | CIVIL = 0
| | | | | | FEMININA = 1: C207
| | | | | | FEMININA = 0
| | | | | | | MOTO = 1: E112
| | | | | | | MOTO = 0: C222
| | | CORPO = 0
| | | | HOSPITAL = 1
| | | | | CIVIL = 1: E111

| | | | | CIVIL = 0
| | | | | | POPULARES = 1
| | | | | | | AGENTE = 1
| | | | | | | | GUARNICAO = 1: C104
| | | | | | | | GUARNICAO = 0: C222
| | | | | | | AGENTE = 0
| | | | | | | | ALVEJADO = 1: C104
| | | | | | | | ALVEJADO = 0: D316
| | | | | | POPULARES = 0
| | | | | | | BAIRRO = 1: C104
| | | | | | | BAIRRO = 0: C222
| | | | HOSPITAL = 0
| | | | | FEMININA = 1
| | | | | | BAIRRO = 1
| | | | | | | ARMA = 1
| | | | | | | | AGENTE = 1: C215
| | | | | | | | AGENTE = 0
| | | | | | | | | FOGO = 1: C207
| | | | | | | | | FOGO = 0: C221
| | | | | | | ARMA = 0
| | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | ESTABELECIMENTO = 0: C112
| | | | | | BAIRRO = 0
| | | | | | | ARMA = 1: C222
| | | | | | | ARMA = 0
| | | | | | | | TOMADO = 1: C221
| | | | | | | | TOMADO = 0
| | | | | | | | | ESTABELECIMENTO = 1: C221
| | | | | | | | | ESTABELECIMENTO = 0
| | | | | | | | | | POPULARES = 1: C222
| | | | | | | | | | POPULARES = 0: C222
| | | | | FEMININA = 0

| | | | | | AGREDIDO = 1
| | | | | | | AGENTE = 1: C114
| | | | | | | AGENTE = 0
| | | | | | | | ARMA = 1: C222
| | | | | | | | ARMA = 0
| | | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | | ESTABELECIMENTO = 0
| | | | | | | | | | GUARNICAO = 1: C222
| | | | | | | | | | GUARNICAO = 0: C222
| | | | | | AGREDIDO = 0
| | | | | | | FOGO = 1
| | | | | | | | SANGUE = 1: C215
| | | | | | | | SANGUE = 0
| | | | | | | | | ALVEJADO = 1: C104
| | | | | | | | | ALVEJADO = 0
| | | | | | | | | | MOTO = 1
| | | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | | POSTO = 0
| | | | | | | | | | | | AGENTE = 1: C222
| | | | | | | | | | | | AGENTE = 0: C222
| | | | | | | | | | MOTO = 0
| | | | | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | | | | ESTABELECIMENTO = 0
| | | | | | | | | | | | AGENTE = 1: C222
| | | | | | | | | | | | AGENTE = 0
| | | | | | | | | | | | | MORRO = 1: C222
| | | | | | | | | | | | | MORRO = 0
| | | | | | | | | | | | | | BAIRRO = 1: C222
| | | | | | | | | | | | | | BAIRRO = 0: C222
| | | | | | FOGO = 0
| | | | | | | ARMA = 1
| | | | | | | | TOMADO = 1

| | | | | | | | | AGENTE = 1
| | | | | | | | | | FATAL = 1: C222
| | | | | | | | | | FATAL = 0: C221
| | | | | | | | | AGENTE = 0: C222
| | | | | | | | TOMADO = 0
| | | | | | | | | BAIRRO = 1
| | | | | | | | | | AGENTE = 1
| | | | | | | | | | | MOTO = 1: C222
| | | | | | | | | | | MOTO = 0: C221
| | | | | | | | | AGENTE = 0: C222
| | | | | | | | BAIRRO = 0: C222
| | | | | | | ARMA = 0
| | | | | | | | POPULARES = 1
| | | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | | ESTABELECIMENTO = 0
| | | | | | | | | SANGUE = 1: C222
| | | | | | | | | SANGUE = 0: C222
| | | | | | | | POPULARES = 0
| | | | | | | | | AGENTE = 1
| | | | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | | | ESTABELECIMENTO = 0
| | | | | | | | | | | MOTO = 1: C222
| | | | | | | | | | | MOTO = 0
| | | | | | | | | | | | HOTEL = 1: C222
| | | | | | | | | | | | HOTEL = 0
| | | | | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | | | | POSTO = 0: C222
| | | | | | | | | AGENTE = 0
| | | | | | | | | | MOTO = 1: C222
| | | | | | | | | | MOTO = 0
| | | | | | | | | | | ESTABELECIMENTO = 1: C222
| | | | | | | | | | | ESTABELECIMENTO = 0

| | | | | | | | | | | | | POSTO = 1: C222
| | | | | | | | | | | | | POSTO = 0
| | | | | | | | | | | | | | BAIRRO = 1: C222
| | | | | | | | | | | | | | BAIRRO = 0
| | | | | | | | | | | | | | | CIVIL = 1: C222
| | | | | | | | | | | | | | | CIVIL = 0
| | | | | | | | | | | | | | | | GUARNICAO = 1: C222
| | | | | | | | | | | | | | | | GUARNICAO = 0
| | | | | | | | | | | | | | | | | MORRO = 1: C222
| | | | | | | | | | | | | | | | | MORRO = 0
| | | | | | | | | | | | | | | | | | RESIDENCIA = 1: C222
| | | | | | | | | | | | | | | | | | RESIDENCIA = 0: C222